INTEGRATION AND MISSING DATA HANDLING IN MULTIPLE OMICS STUDIES

by

Zhou Fang

MS in Industrial Engineering, University of Pittsburgh, 2014

BS in Theoretical and Applied Mechanics, Fudan University, China, 2012

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate school of Public health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF BIOSTATISTICS

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Zhou Fang

It was defended on

May 3rd 2018

and approved by

George C. Tseng, ScD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Wei Chen, PhD, Associate Professor, Department of Pediatrics, School of Medicine, University of Pittsburgh

Gong Tang, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Ming Hu, PhD, Assistant Staff, Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation

Ying Ding, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Advisors: **George C. Tseng**, ScD, Professor, Department of Biostatistics,

Graduate School of Public Health, University of Pittsburgh,

Wei Chen, PhD, Associate Professor, Department of Pediatrics, School of Medicine,

University of Pittsburgh

INTEGRATION AND MISSING DATA HANDLING IN MULTIPLE OMICS STUDIES

Zhou Fang, PhD

University of Pittsburgh, 2018

ABSTRACT

In modern multiple omics high-throughput data analysis, data integration and missingness data handling are common problems in discovering regulatory mechanisms associated with complex diseases and boosting power and accuracy. Moreover, in genotyping problem, the integration of linkage disequilibrium (LD) and identity-by-descent (IBD) information becomes essential to reach universal superior performance. In pathway analysis, when multiple studies of different conditions are jointly analyzed, simultaneous discovery of differential and consensual pathways is valuable for knowledge discovery. This dissertation focuses on the development of a Bayesian multi-omics data integration model with missingness handling, a novel genotype imputation method incorporating both LD and IBD information, and a comparative pathway analysis integration method.

In the first paper of this dissertation, inspired by the popular Integrative Bayesian Analysis of Genomics data (iBAG), we propose a full Bayesian model that allows incorporation of samples with missing omics data as well as a self-learning cross-validation (CV) decision scheme. Simulations and a real application on child asthma dataset demonstrate superior performance of the CV decision scheme when various types of missing mechanisms are evaluated.

In the second paper, we propose a genotype inference method, namely LDIV, to integrate both LD and IBD information. Both simulation study with different family structures and real data results showed that LDIV greatly increases the genotype accuracy, especially when family structures are informative.

The third paper presents a meta-analytic integration tool, Comparative Pathway Integrator (CPI), to discover consensual and differential enrichment patterns, reduce pathway redundancy, and assist explanation of the pathway clusters with novel text mining algorithm. We applied CPI to jointly analyze six psychiatric disorder transcriptomic studies to demonstrate its effectiveness, and found functions confirmed by previous biological studies and new enrichment patterns.

All three projects could have substantial public health importance. By handling missing data, a higher statistical power and accuracy in clinical prediction and biomarker selection can be retained by FBM given fixed budget and sample size. LDIV effectively increases genotyping accuracy. CPI simultaneously discovers biological processes that function differentially and consensually across studies. It will also assist scientists to explore pathway findings with reduced redundancy and more statistical backup.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The purpose of this dissertation is to propose new methods in genomics data analysis when information could be integrated but is susceptible to missingness. The original passion of this topic arises from the rapid evolving of high-throughput genomic data generating technologies, as well as the need of methods to approach the underlying truth behind the accumulating data. The methods proposed in this thesis include a Bayesian hierarchical model for multi-omics data with missingness, a genotype inference methods incorporating both LD and family information, and a pathway integrative analysis framework to capture differential and consensual pathway enrichment patterns. These methods are supposed to be of interests to both scientist looking for methods that could facilitate knowledge discovery or increase accuracy or power of the finding, and biostatisticians seeking a base to further extend the performance.

These research work could never been complete without the wonderful mentorship I received during my four years of PhD studies. I would like to thank Dr. George Tseng for his detailed instructions and guidance, unfailing support and encouragement, and himself as an excellent example of being a knowledgeable statistician who humbly serves the scientific community. His help is not confined to my studies and research, but also to me as a person. Sometimes I sense that he knows me better than I know myself. I also want to give my gratitude to Dr. Wei Chen, who has offered his most generous time in training and correcting me, and preparing me as a bridge between mathematics and clinical research. He is the person who led me to bioinformatics field, and opened up a new world for me on how statistical computing can help clinical research and in turn, saving lives.

I also want to thank my committee members Drs. Gong Tang, Ming Hu, and Ying Ding, who have put tremendous efforts and knowledge into the research works in this dissertation. I

am also greatly influenced and encouraged by their passions and positive attitudes in science. I also want to thank all my lab mates for the four years of precious time we shared. They are indispensable part in an ecosystem for academic excellence. And I have been touched by their perseverance.

I would like to express my deepest gratitude to my father, Ming, and my mother, Ying-ping, who have loved and taught me how to love for the best twenty-eight years a son can ever ask for. The trustworthiness I saw on my father and the wisdom I saw on my mother have always been the inspiration and drive of my life.

Lastly, I want to thank my love and my wonderful wife, Xiaoying. You have been a real blessing to me. We think alike. We read each other's minds. And I am positive that we would walk along this path, thinking in the end, how lucky we were keeping each other's company.

1.0 INTRODUCTION

1.1 OVERVIEW OF HIGH-THROUGHPUT OMICS DATA AND TECHNOLOGIES

Omics refers to the studies of a group of fields ending with -*omics* in biology, such as genomics, transcriptomics, epigenomics, proteomics, and etc. PubMed and Google search entries sees an explosive increase in the use of terms related to omics and omes starting from mid '90s (http://bioinfo.mbb.yale.edu/what-is-it/omes/omes.html). Omics data are essential components to the central dogma in molecular biology ($DNA \leftrightarrow RNA \rightarrow Protein$) (Watson, 1965). With the emerging technologies to generate high-throughput omics data and the methods developed to conduct in-depth analysis on those data, scientists are now gaining insights into biology processes, and bring to light their associations in pathology. This section will briefly introduce a few omics data types relevant to this thesis, two major technologies in omics data generation, and public omics databases.

1.1.1 Genomic data

Genomics data analysis studies information carried by the genomes of organisms. Genomes are usually consist of DNA (Deoxyribonucleic acid), but are sometimes RNA (Ribonucleic acid) in RNA viruses. DNA has two strands, stabilizing the storage of genetic material. Each strand is composed of simpler units named nucleotides of four types, cytosine (C), guanine (G), adenine (A) or thymine (T). The different combinations of the nucleotide on the genome translate to the diversity of species and organisms within species. The human genome contains around 3×10^9 base pairs nucleotide, distributed among 22 paired chromosomes and

two sex chromosomes. Only a small fraction of them (approximately 1.5%) (Lesk, 2017) are protein-coding sequences, meaning they could to be translated and coded into proteins. And the other parts function as regulatory sequences, associated as non-coding RNA molecules, and etc.

Although for human beings, the differences among genomes of individuals are only on the order of 0.1%, this difference explains a great amount of differences in our phenotypes, or sometimes diseases. Genetic variations are fundamentally brought about by mutations, which are permanent alterations of nucleotides in genetic materials. Mutations will be carried through when DNA replicates itself, or be transcribed to RNA, and thus might be expressed as different protein coding. If the mutation happens in somatic cells and thus only takes effect for the individual with the mutation but not passed down to offspring, it is called a somatic mutation, which is usually seen in tumor cells. When mutation happens in reproductive cells like sperm, it is inheritable, and is called a germ line mutation. Chromosome crossovers and recombination during meiosis can also bring in genetic variations.

There are various types of genetic variations. Single nucleotide polymorphism (SNP) is the most common type of genetic variation, it indicates the variation in a single nucleotide. It has been discovered that over thousands of single nucleotides are potentially associated with certain phenotypes or diseases through Genome-wide association studies (GWAS) (Hindorff et al., 2009; McCarthy et al., 2008).

Another type of genetic variation is copy number variation (CNV), a structural variation with deletion or duplication of a large chromosome region. It has been found that CNV is associated with disease phenotypes, gene expression regulation, and other genomic processes (McCarroll and Altshuler, 2007). Other genetic variation includes insertion/deletion (indel) polymorphism, which represents the addition/missing of a certain nucleotide sequences.

1.1.2 Transcriptomic data

Transcriptomics studies transcriptome, the pool of RNA molecules. RNA is essential in coding, regulation and expression of genes. Unlike DNA, RNA is single-stranded molecule, and does not have thymine, but have uracil (G) instead as their bases. There are various types

of RNA. For a certain gene to be expressed and coded into protein, messenger RNA (mRNA) is necessary, which is the RNA that conveys information from the genome to ribosome to instruct the coding of proteins. This step is called transcription. In this step, mRNAs from the same gene can have various combinations of exons of this gene, and thus produce different protein isoform in later steps; this phenomenon is called alternative splicing. When mRNA is in ribosome, transfer RNA (tRNA) transfers amino acids to ribosome, and a catalytic component of ribosomes, the ribosomal RNA (rRNA), links amino acids to synthesize a specific protein. This step is called translation. MicroRNA (miRNA) is a short RNA found in eukaryotes that, along with other types of regulatory RNAs, can down-regulate gene expression by binding with a part of mRNA or DNA to prevent transcription.

Expression quantitative trait loci (eQTLs) are genomic loci that affect mRNA expression levels. The eQTLs that are mapped approximately to their original gene location is called local eQTLs, or *cic* eQTLs; on contrary, if it is mapped far from original gene location, or even on other chromosomes, it is referred to as distant eQTLs, or *trans* eQTLs. Lappalainen et al. (2013) investigated mRNA, miRNA, together with genetic variations of 462 individuals from the 1000 Genomes projects, and discovered different types of genetic variations affecting regulation of most gene in both transcript structures and expression level.

1.1.3 Epigenetic data

Epigenetics studies a pool of epigenetic modifications, reversible modifications that would not alter DNA sequences on organism's genetic material. DNA methylation and histone modification are two most commonly studied epigenetic modifications. DNA methylation happens when a methyl group is added to cytosine in DNA sequences. And it will repress the gene expression by reducing accessibility of DNA sequence with methylated cytosine to transcriptional machinery, for example, RNA polymerase. DNA methylation is generally related to imprinting, x chromosome inactivation and silencing of repetitive DNA (Schübeler, 2015), and for vertebrates, heritable methylation only occurs at CpG dinucleotide, and approximately 60% of CpG sites are methylated to DNA regions with many repetitive CpG sites, i.e. CpG island (Bird, 2002). Two values are usually used to measure methylation

level, β value and m value. β value is continuous between 0 and 1, indicating the percentage of methylation events among all cells, and is widely accepted for its interpretability. On the other hand, m value fits better with Gaussian distribution, and is more valid for statistical modeling and analysis (Du, 2010). Each of the two values can be transformed to the other with the following function:

$$M = log_2(\beta/(1-\beta))$$

Histone is the most common type of chromatin, the protein-DNA complexes that condenses DNA sequences into a 3D coiled structure. Histone modification regulates gene expression by either disrupting contact between nucleosomes, or nucleosome remodeling.

1.1.4 High-throughput technologies in omics research

1.1.4.1 Microarray DNA microarray technology emerged in early 1990s, and marked the beginning of high-throughput omics data generation and analysis era. The older technologies that are therefore replaced by microarray, for example, northern blot, are time consuming, expensive, and limits gene detection to order of tens, while microarray enables the detection of thousands of genes or even whole genome gene expression profiling in reasonable time with a feasible budget. Microarray is made up with array of tens of thousands of transcript-specific probes on microscope glass slide or silicon chips. If RNA rather than DNA is to be detected, the RNA samples will first need to be reversely transcribed to cDNA. Then the DNA is fluorophore-, silver-, or chemiluminescence-labeled to be scanned later. Then they are hybridized to microarray chips and washed away, while the fluorescence will be captured by machine with a certain intensity, which corresponds to the gene expression level. Other than gene expression level profiling, there are other types of microarray chips, for example, SNP array, for SNP genotyping detection, or copy number variation.

However, there are some limitations of microarray technology. Firstly, it measures the relative abundance instead of direct count as later technologies like RNASeq, and the signal is linear only over a certain range of concentration. Secondly, designing array that can be distinguishable for all genes and all alternative splicing of genes is too difficult. Finally, microarray cannot detect sequences we have no prior knowledge of, i.e., each array can

only capture the specific sequence it is designed for. With those limitations and the rapid development of new sequencing technologies, microarray becomes less popular, yet is still competitive in terms of its cost.

1.1.4.2 Next generation sequencing The next-generation sequencing (NGS) technology emerged with the increasing demand for inexpensive, accurate and fast generation of high-throughput omics data (Metzker, 2010). The name next-generation sequencing comes from its predecessor, automated Sangers sequencing. The development of NGS greatly sped up scientific findings, and urged statisticians to rethink their methods to apply to the accumulating big data in omics. There are a wide variety of fields NGS has been applied to, including whole genome sequencing, exome sequencing, RNA transcriptome profiling (RNA-seq), bisulfite sequencing for DNA methylation, DNA-protein interaction (ChIP-seq), etc. NGS is also called shotgun sequencing, due to its unique processing steps: first, the whole sequence (DNA or cDNA) will be broke down to shorter fragments, which are called reads, then with PCR amplification technique, each read will have thousands of copies to form a cluster. The machine will then scan the amplified cluster, and sequencing data is generated. Multiple platforms are available for this technology, for example, Roche, SOLiD, Illumina, and so on. After reads are sequenced, algorithms are developed to assemble (*de-novo* assemble) the reads into a whole sequence, or to align (assemble against reference) reads against a reference genome. The count of reads mapped to a certain gene is called depth, or coverage, which serves an important indicator for the intensity of the gene expression, or the quality and confidence of a certain assembly. The count data are less commonly used compared to continuous data in statistical modeling, so the count of reads could then be transformed to continuous value such as TPM (Transcripts Per Kilobase Million), RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million). In the chapter 2 of this thesis, we will be mainly using TPM to quantify RNA-seq gene expression level.

NGS has several advantages compared to microarray. First, it can identify the transcripts or sequences not designed or known in hybridization-based approaches, thus it enables many novel discoveries, for instance, detecting de novo mutation. Secondly, NGS has very low background signal, greatly increasing the accuracy in both high and low intensity end. Moreover, NGS also detects larger dynamic range for expression level (Wang et al., 2009).

1.1.5 Public omics data repositories

Massive data of various omes has been accumulated over the past two decades, due to the breakthrough advances of high-throughput omics data generating technology. The amount of data generated, stored and analyzed can foreseeably increase even faster with the dropping in cost of sequencing technology, and the increase of data storage space and computational power by Moore's law. Therefore, public omics data repositories play an essential role in between data-generating labs, scientists, and statisticians. Starting with genomic data, there are Encyclopedia of DNA Elements (ENCODE, The ENCODE Project Consortium (2012)) and its extension, MODel organism ENCyclopedia Of DNA Elements (modENCODE, Gerstein et al. (2010)) project serving to identify and annotate functional elements of human genome. The 1000 Genome Project (IGSR, The 1000 Genomes Project Consortium (2015)) aims to find most relatively common genetic variants with minor allele frequency greater than 1%. One popular kind of studies to find the association between phenotype and human genome is the genome-wide association studies (GWAS). Hundreds of GWAS (GWAS Catalog, www.ebi.ac.uk/gwas) are conducted in the past decades, and stored in Genotypes and Phenotypes databases (dbGaP). For transcriptomic data, there are Gene Expression Omnibus (GEO, Edgar et al. (2002)), Sequence Read Archive (SRA, Leinonen et al. (2011)), and Array Express (www.ebi.ac.uk/arrayexpress). There are also worldwide or nationwide consortium focusing on collecting omics data within a certain domain, one of the largest is The Cancer Genome Atlas (TCGA, Weinstein et al. (2013)) initiated by National Caner Institute. TCGA has a wide collection of a total of 33 cancer types with DNA mutation, CNV, RNA gene expression, miRNA expression, protein expression, DNA methylation, etc.

The numerous types of omics data publicly available provided both opportunities and challenges in a statistical domain named vertical omics data integration. This concept will be introduced in next subsection.

1.2 VERTICAL OMICS DATA INTEGRATION

With the accumulation of high throughput multiple omics data from numerous labs, a domain named omics data integrative analysis arose. There are two types of omics data integration, the horizontal integration, and the vertical integration. The former one combines studies/cohorts with the same omics data types, e.g., transcriptomic studies, and aims on the gain of statistical power and reproducibility by integrating them. The latter one is relative to Chapter 2; it combines multi-omics data from the same cohort of individuals, investigates the inner regulatory mechanisms in between omics data (e.g. gene expression, eQTL, and DNA methylation), and seeks to gain deeper insight into the corresponding biological processes. A few successful examples in vertical meta-analysis are ovarian cancer (Cancer Genome Atlas Research Network, 2011), breast cancer (Cancer Genome Atlas Network, 2012), and stomach cancer (Cancer Genome Atlas Research Network, 2014) from TCGA. There are also vertical integration methods developed for different purposes. For example, for the purpose of prediction on a clinical outcome yet being able to uncover the regulatory mechanism between omics data types, Wang et al. (2012) proposed integrative Bayesian analysis of genomics data (iBAG). And for the purpose of clustering, a few examples are: Bayesian consensus clustering (BCC) fitting a finite Dirichlet mixture model, allowing for both common and omic-specific clustering patterns (Lock and Dunson, 2013); iCluster used matrix decomposition and regularization in latent variables to help clustering (Shen et al., 2009a), sparse k-means clustering allows clustering with overlapping feature groups (Huo and Tseng, 2017a).

In Chapter 2, we focus on iBAG model for vertical omics integration, with the aim of prediction and feature selection. Our model is able to discover inner regulatory mechanisms between different omics data types (in our example, RNAseq gene expression and DNA methylation), and our model also incorporate missing value handling, which will be introduced in the next subsection.

1.3 MISSING DATA IMPUTATION IN OMICS STUDIES

Missing data can exist in almost all types of statistical analyses. One way to categorize missing data is by a typology by Rubin (1976), or the three missing assumptions. First one is missing complete at random (MAR). Assuming a generic notation, where R are the probabilities of missingness, and data Y are partitioned to two part, $Y_{com} = (Y_{obs}, Y_{mis})$, and that:

$$P(R|Y_{com}) = P(R|Y_{obs})$$

This means the distibution of missingness does not depend on missing data Y_{mis}. A special case of MAR is missing complete at random (MCAR), which further assume missing probability does not depend on Y_{obs}, i.e.

$$P(R|Y_{com}) = P(R)$$

MAR and MCAR together are called ignorable missingness, as in these cases one can ignore the model of missingness, i.e. no need to model R. Yet the missing data itself is still essential and usually need to be modeled.

A third missing assumption is called missing not at random (MNAR), which happens when MAR assumption is violated. MNAR is also called non-ignorable missingness, and is substantially harder to deal with when modeling (Schafer and Graham, 2002). Unfortunately, in real data, there is no way to examine those assumptions, because those missing data per se could not be observed.

Another way to categorize missing data is by exchangeability, where *exchangeable* means the distribution of missing data (not R, the missingness) does not depend on other observed data, and that *nonexchangeable* means that the distribution of missing data depend on other observed data.

Same as other statistical problem, missing data is inevitable in omics data analysis. This can be due to various technical limitations (Aittokallio, 2010; Albrecht et al., 2010). Those types of missing usually come with smaller percentage (e.g. less than 10%), and numerous methods have been developed to address it. However, in vertical integrative analysis of omics data, another type of missing will occur by the mismatch of data types of samples. Since in

integration, it is likely that not all samples are measured with all omics data, the sample size will be drastically reduced when taking intersection across all omics data (Lin et al., 2016). And the missing proportion for this missing type could usually be more than 30%, and will be a more severe hindrance to the integration than the previous type. Therefore, chapter 2 will mainly deal with this type of missing.

Numerous methods have been developed to deal with missing data in general as well as domain-specific. Without imputation, complete case (CC) analysis is an unbiased method under MAR assumption, which is the approach by simply taking the cases with complete data for further analysis. Yet the power decrease is worth noting especially when missing proportion is large. Among imputation methods, a few of the most generic ones are: mean substitution, hot deck, conditional mean, and predictive distributions (Schafer and Graham, 2002). Multiple imputations (MI), impute by generating $K > 1$ imputed values for missing observations from appropriate probability distribution, are proven to be a robust substitution for aforementioned imputation methods (Rubin, 1987), and is gaining its popularity with the development of MCMC approaches. Lastly, in full Bayesian model, missing data are treated the same as parameters, and therefore need to be modeled, and then estimated or sampled, in a similar fashion. There are also imputation methods specifically developed for missing problem in omics data. For example, Voillet et al. (2016) applied multiple imputations to multiple factor analysis; Lin et al. (2016) imputes missing data by taking into account the correlation across different omics data in vertical integration.

In Chapter 2, we will propose our full Bayesian framework with missing data imputation modeled together with parameter estimation and feature selection.

1.4 BAYESIAN MODELING FOR PREDICTION AND FEATURE SELECTION

Bayesian inference solves statistical problems by building probability model to data, using parameters of the model to generate probability distributions, and hence summarizing the results (Gelman et al., 2014). Different from frequentist inference, Bayesian inference assumes true parameters are from certain distributions with variance, instead of fixed. A standard

Bayesian process could be summarized as three steps:

- Setting up *full probability model*, which is a joint probability distribution of all observed and missing data.
- Calculate the appropriate *posterior distribution* conditioning on observed data. This step follows the Bayesian rule that $P(M|D) \propto P(M) \times P(D|M)$, where D means the fully observed data, M means the full parameter set, and thus $P(M|D)$ is the posterior distribution, $P(M)$ is the prior distribution, and $P(D|M)$ is the likelihood.
- Evaluate the model by checking model fitting, if conclusions are reasonable, and conducting sensitivity analysis.

Since mathematically, step 2 requires taking expectation of integrals to derive the posterior distribution, it is usually hard or even impossible to find analytical solution to this problem. Therefore, although Bayesian inference emerged over two hundred years ago, it is not popular until the development of Monte Carlo Markov chain (MCMC) algorithms and the booming of modern computation technologies. MCMC is by nature a random walk algorithm, with idea of stochastically sampling parameter in its parameter space, while keeping Markov property, i.e. the status in current iteration, conditional on one previous iteration, is independent on all earlier iterations (Gelman et al., 2014). Metropolis-Hastings (MH) is one of the most famous and conventional methods for MCMC. It first generates proposal distribution, then by accept-reject criteria to decide the distribution to sample from in the next iteration (Metropolis et al., 1953; Hastings, 1970). One typically applied special case of MH is Gibbs sampling (Geman and Geman, 1984), where the conditional distribution is used as sampling distribution.

Bayesian approaches are gaining increasing popularity in many statistical domain, including omics data analysis. This is due to several advantages of it compared to frequentist inference. First, Bayesian inference allows taking prior distribution of parameter into the model, thus is suitable for omics data analysis where some prior knowledge are obtained from scientific experiments. Secondly, Bayesian inference allows establishing of relatively complex models in multiple layers, called *Bayesian hierarchical model*, with no substantial modeling difficulty added. The model flexibility makes it suitable for addressing correlation

relationship between multiple omics data types and outcomes. Another merit of Bayesian inference is its ease of interpretation for results, i.e. usually probability is given instead of p-value in frequentist analysis.

An increasing amount of Bayesian methods have been developed for omics studies. Many of them aim at prediction and feature selections, for example, iBAG (Wang et al., 2012) use Laplacian priors (Park and Casella, 2008) to put shrinkage on parameters for feature selection, and therefore detect important genes or biomarkers associated with clinical outcome; Ishwaran and Rao (2005) use spike-and-slab prior (George and McCulloch, 1993) to identify biologically important signals while minimizing false discoveries in multi-group microarray data.

In Chapter 2, we will discuss in detail a full Bayesian model to integrate multiple omics data types for clinical outcome prediction, feature selections, and full Bayesian missing data imputation.

1.5 STATISTICAL INFERENCES IN GENOTYPE CALLING

The magnitude of genetic variants confidently discovered by next generation sequencing (NGS) technologies has empowered a variety of biologically meaningful downstream analysis. For instance, genome-wide association studies (GWAS) (Genomes Project Consortium et al., 2012) have revealed significant association between thousands of genetic variants with traits or diseases. The majority of those variants are single nucleotide polymorphisms (SNPs). A first and fundamental step of the NGS downstream analysis is called genotype calling, which is the step for inference of the genotypes of those SNPs based on raw sequencing reads. An accurate genotype calling is crucial and the prerequisite for any further downstream analysis based on SNPs detected. It has been shown that low-depth sequencing, which means the average sequencing read depth on each sites is relatively low (<5X) or moderate (<10X), could be more powerful and cost-efficient, when an appropriately larger sample size is used (Li et al., 2011). Large sequencing projects like 1000 Genome Project, UK10K project and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Project

has been taking this low-depth large-cohort strategy, for example, 1000 Genome Project sequenced 2535 subjects at the depth of 4-6X. Yet due to a non-ignorable per-base error during the sequencing, uncertainty of low-depth genotype calling could be high, therefore sophisticated likelihood-based methods are almost always required for an accurate call.

Existing accurate genotype calling methods developed for low-coverage sequencing scenarios could be divided into two categories, based on the extra constraints and information they consider apart from raw sequencing reads. One category of methods improve genotype calling accuracy by considering the linkage disequilibrium (LD) between SNPs. Hidden Markov models (HMM) are often required in inferring the LD patterns, a few examples are Thunder, MaCH, and Beagle4. There are also methods doing inference using EM algorithm for the same purpose, for example, PedHapGC (Zhou and Whittemore, 2012). The second category of methods are developed as family-based sequencing projects emerged and showed its strengths in detecting rare SNPs, studying parent-offspring origins, etc. These methods take pedigree structures, presented in identity-by-descent (IBD) information, into consideration. In this category, PedHapGC uses EM-algorithm-based approach to account for pedigree constraints; more other methods are based on HMM. For example, TrioCaller (Chen et al., 2013) is developed to address parents-offspring trio family structure; FamLDCaller (Chang et al., 2016) further considered general family structure by looping through each parent-offspring trio within a family; Polymutt2 (Li et al., 2015) uses inheritance vector to represent IBD pattern, which considered not only the parent-offspring Mendelian constraints, but also the information shared between siblings; and lastly, the methods we propose in chapter 3 of this thesis, namely LDIV, leverages both LD and pedigree information for jointly inference of the genotypes. Other than these two main categories, some other popular methods are GATK (McKenna et al., 2010) and SamTools (Li et al., 2009), which takes neither of the aforementioned information. Though being useful for preliminary calls, simulation and real data analysis shows that these methods are less powerful in the preferable low-coverage large-cohort sequencing scenarios.

In Chapter 3, we propose our statistical methods named LDIV to address the above issue. And we benchmark the performance of each method in different sequencing scenarios, with varying sequencing depths, per-base error rate, and family structure.

1.6 PATHWAY ENRICHMENT ANALYSIS

In a typical transcriptomic study, a set of candidate genes associated with diseases or other outcomes are first identified through differential expression analysis. Then, to gain more insight into the underlying biological mechanism, pathway analysis (a.k.a. gene set analysis) is usually applied to pursue the functional annotation of the candidate biomarker list. The rationale behind pathway analysis is to determine whether the detected biomarkers are enriched in pre-defined biological functional domains. These functional domains might come from one of the publicly available databases such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). Three main categories of pathway analysis methods have been developed in the past decade. The first method called "over-representation analysis" considers biomarkers at a certain DE evidence cutoff and statistically evaluate the fraction of DE genes in a particular pathway found among the background genes. Without a hard threshold, the second category "functional class scoring" takes the DE evidence scores of all genes in a pathway into account and aggregates them into a single pathway-specific statistics. The third category "pathway topology" further incorporates the information of inter-gene interaction and their cellular location in addition to the pathway database (Khatri and Butte, 2012).

Pathway analysis methodology for single study is a relatively mature field (Huang et al., 2009), yet methods for pathway analysis in integrative studies is lacking (Shen and Tseng, 2010). A meta-analytic pathway analysis approach that allows the discovery of both consensual and differential pathway enrichment patterns will be introduced in chapter 5.

1.7 OVERVIEW OF THE DISSERTATION

My disseration contains five chapters. Chapter 1 contains introduction of high-throughput omics data and technologies, statistical methods motivated by the integration and missing data in multiple omics data analysis, basic knowledge in Bayeisan data analysis, genotype calling methods, and pathway enrichment analysis methods. This chapter serves as the back-

ground knowledge and motivation for methodology development and application in Chapter 2, 3 and 4.

Chapter 2 introduced a full Bayesian hierarchical model for vertical multi-omics data integration, with missing data handling. Inspired by iBAG model, our model has non-linear feature selection empowered by spike-and-slab prior, and use three layers of Bayesian model to incorporate data integration and imputation simultaneously. In addition, we developed a cross-validation-based approach to facilitate user choosing best strategy when facing missing data in vertical integration. The content in this Chapter is under review at Bioinformatics.

Chapter 3 introduced a HMM-based model, namely, LDIV (linkage disequilibrium and inheritance vector), to improve genotype calling in family-based sequencing data. In this model, both linkage disequilibrium (LD) patterns and inheritance by descent (IBD) pedigree constraints are comprehensively incorporated, and the genotype likelihoods are jointly inferred in a family. This method is both accurate and computationally feasible.

Chapter 4 introduced a framework of meta-analytical integration of multiple transcriptomic studies for consensual and differential pathway analysis, wrapped in a tool named Comparative Pathway Integrator (CPI). Our tool incorporates 24 pathway databases, and is able to identify both commonly and study-specifically enriched pathways, empowered by adaptively weighted Fisher's method. This tool also reduces pathway redundancy analysis by clustering analysis based on the overlapping genes among enriched pathways, and automates the annotation of pathway clusters by text mining, offering more statistically valid summarization. CPI also provides users both text and graphical outputs for intuitive while statistical solid presentation and easy interpretation. The content in this chapter is under preparation.

Chapter 5 is the discussions and future works. We will apply our LDIV methods to real studies. We also propose to do a comprehensive comparison and review of all current genotype calling methods to give the scientific community an informative guideline.

2.0 A BAYESIAN MODEL FOR INTEGRATING HIGH-THROUGHPUT MULTI-OMICS DATA WITH MISSINGNESS HANDLING

2.1 INTRODUCTION

Multi-level omics data refer to the combined molecular data in various types, for example genome, transcriptome, methylome and proteome data, measured on a common cohort of patients. The availability of multi-level omics data in both magnitude and varieties poses challenges as well as opportunities to understand fundamental mechanisms of diseases and pathologies. Compared to separately discovering the association patterns between each omics data and phenotypes, an integrative framework that simultaneously integrates multiple omics data types will uncover more insightful regulatory machineries between different omics data, and will hence deepen our understanding to hereditary and environmental causes in pathology (Richardson et al., 2016; Tseng et al., 2012, 2015).

In the literature, many strategies have emerged for integration of multi-omics data. To cluster samples for identifying unknown disease subtypes, integrative clustering (iCluster; Shen et al., 2009b), Bayesian consensus clustering (BCC; Lock and Dunson, 2013), group structured integrative clustering (GS-iCluster; Kim et al., 2017) and integrative sparse K-means (IS-Kmeans; Huo and Tseng, 2017b) have been developed for integrative clustering of multi-omics data. For association and prediction modeling, iBAG (Wang et al., 2012) investigates association patterns of mRNA expressions and methylation with clinical outcome via a mechanistic model for associating methylation with gene expression and then a clinical model for direct association between expression and clinical outcome or via methylation. A Bayesian hierarchical model is then established for the inference. Such Bayesian hierarchical modeling has gained popularity in multi-omics integrative analysis due to its flexibility in

model construction for complex regulatory structure, convenience to incorporate prior biological knowledge and advances in modern computing. In the association and prediction modeling, we usually focus on two key goals in the inference: firstly, to select predictive biomarkers for the phenotype and secondly, to predict clinical outcome from the selected biomarkers. As tens of thousands of features are available in omics data, the first goal of feature selection is essential in interrogating the biological and pathological mechanism of a targeted disease. The second goal could be of immediate clinical use, for example, assisting diagnosis of a disease, directing the best treatment decision and predicting drug response.

One obstacle for applying multi-omics integrative methods in real applications is missing data. Due to various reasons(e.g. limited budget, bad tissue quality or insufficient tissue amount), it is common that only partial samples have all omics data types (Voillet et al., 2016). For example, TCGA breast cancer study had 922 samples measured in methylation array, yet only 781 were measured in miRNA expression and 587 were measured for gene expression. Almost 40% of the samples are missing at least one type of omics data. To circumvent this pitfall, a naïve and convenient solution is by complete-case (CC) analysis, where samples with any missing measurement are ignored. This approach results in dramatic decrease of sample size and thus decreases of statistical power, especially when more omics data types are combined. The shortcoming of CC in omics studies is recently noticed by statisticians, for example, Voillet et al. (2016) who developed a multiple imputation approach focusing on multiple factor analysis for multiple omics data. However, a unified framework serving the aforementioned purpose of feature selection, prediction, as well as missingness handling is still lacking. In this paper, we are motivated by the iBAG model combining mRNA expression methylation, clinical variables to predict a targeted continuous outcome. We propose a *full Bayesian model with missingness (FBM)* that allows iBAG to handle situations when partial samples are missing with mRNA expression or methylation. Extensive simulations and real applications demonstrated superior performance in feature selection and prediction accuracy of the new approach compared to naïve complete-case approach. The model can be extended to other omics data types or other targeted outcomes (e.g. binary or survival).

The paper is structured as the following. In Chapter 2.1, we introduce the motivation, the original iBAG model and the proposed FBM. Chapter 2.2 discusses the inference of prediction and feature selection of FBM (Chapter 2.2.1) and evaluation benchmarks (Chapter 2.2.2). Chapter 3 contains extensive simulations to evaluate performance of FBM and Chapter 4 proposes a *cross-validation (CV) decision scheme* to determine whether and how to incorporate samples with missingness in FBM. Chapter 5 includes an application to a childhood asthma dataset with 460 individuals. Final conclusion and discussion are presented in Chapter 6.

2.2 METHODS

2.2.1 Motivation and the full Bayesian model with missingness

2.2.1.1 Motivation IBAG is a two-layer Bayesian hierarchical model for vertical integrative analysis of multi-level omics data, assuming data are complete. However, in reality, a large proportion of missing data is commonly seen due to budget or limitation in tissue collection. Figure 1a gives an example of data structure with missingness. Suppose there are a total of N samples, $\mathbf{Y}$ indicates the clinical outcome of interest; C indicates clinical factors; $\mathbf{G}_{N\times K} = (\mathbf{G}'_{obs,N^G_{obs}\times K}, \mathbf{G}'_{mis,N^G_{mis}\times K})'$ indicates the gene expression with missingness, where $\mathbf{U}^G = (U^G_1, ...U^G_N)$ is the missing indicator; likewise, $\mathbf{M}_{N\times J} = (\mathbf{M}'_{obs,N^M_{obs}\times J}, \mathbf{M}'_{mis,N^M_{mis}\times J})'$ indicates methylation data with missingness, where $\mathbf{U}^M = (U^M_1, ...U^M_N)$ is the missing indicator. $U = 1$ indicates missing and $U = 0$ indicates observed. For example, number of samples with methylation level missing is $N^G_{obs} = N - \sum_{k=1}^{K} U^G_k$. We assume $U^M_i \times U^G_i \neq 1$ for $\forall\ 1 \leq i \leq N$. The original iBag model takes only the complete data and is subject to loss of statistical power. To handle missingness, we propose a full Bayesian model with missingness inspired by iBAG model, to achieve three goals simultaneously: feature selection, prediction, and missing data imputation. In Chapter 2.1.2, we briefly introduce the iBAG model. In Chapter 2.1.3, we propose our full Bayesian model for multi-omics integration with missingness imputation.

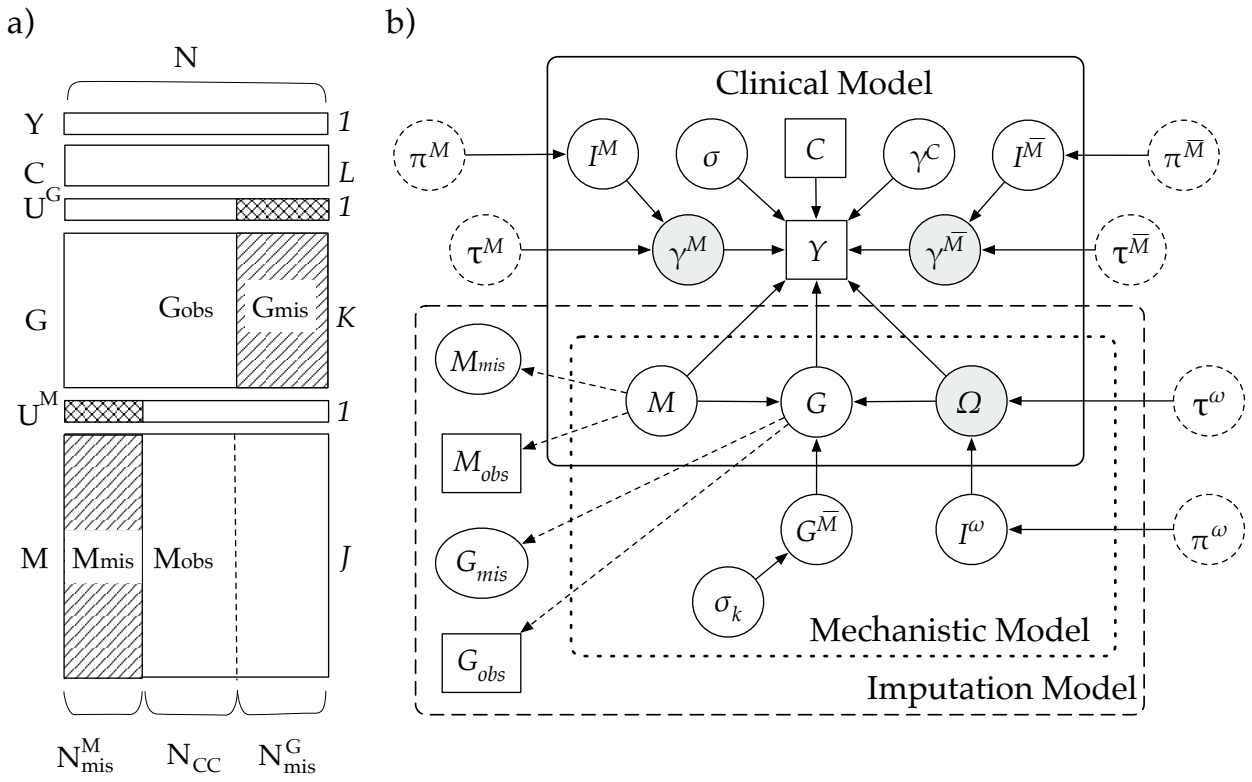


Figure 1: Illustration of missing pattern and model parameters.

a) Illustration of missing pattern adopted in this paper. Slash-shaded area represents missing data, cross-shaded area represents value 1 for missing indicator vectors U^G and U^M. b) DAG of the model and parameters. The square in the DAG denotes observed data, solid circle denotes variable to be updated, circle in grey color is the variable of interests, dashed circle denotes prior. Solid arrows indicate stochastic dependencies, and dashed arrows indicate deterministic dependencies.

2.2.1.2 iBAG In iBAG, the mechanistic models in the first layer assess gene-methylation effect and divides gene expressions into two parts, the part regulated by methylation ($\mathbf{G}^M$) and the part regulated by other mechanisms ($\mathbf{G}^{\overline{M}}$).

$$\mathbf{G} = \mathbf{G}^M + \mathbf{G}^{\overline{M}}, \mathbf{G}^M = \mathbf{M}\mathbf{\Omega},$$

where $\mathbf{G}^M = (g^M_{nk})_{N\times K} = (\mathbf{g}^M_1, ..., \mathbf{g}^M_K)$, $\mathbf{G}^{\overline{M}} = (g^{\overline{M}}_{nk})_{N\times K} = (\mathbf{g}^{\overline{M}}_1, ..., \mathbf{g}^{\overline{M}}_K)$, $\mathbf{\Omega} = (\omega_{jk})_{J\times K} : \omega_{jk}$ denotes the "gene-methylation" effect that estimates the (conditional) effect of jth methylation feature on the kth gene.

The second layer called clinical model assesses the association between gene expression ($\mathbf{G}^M$ and $\mathbf{G}^{\overline{M}}$) and the phenotype:

$$\mathbf{Y} = \mathbf{C}\gamma^C + \mathbf{G}^M\gamma^M + \mathbf{G}^{\overline{M}}\gamma^{\overline{M}} + \epsilon,$$

where $\gamma^C = (\gamma^C_1, ...\gamma^C_L) : \gamma^C_l$ denotes the effect of the lth clinical factor on the clinical outcome Y. $\gamma^M = (\gamma^M_1, ...\gamma^M_K) : \gamma^M_k$ denotes the gene expression effect of $\mathbf{g}^M_k$on clinical outcome Y, called a type M effect. $\gamma^{\overline{M}} = (\gamma^{\overline{M}}_1, ...\gamma^{\overline{M}}_K) : \gamma^{\overline{M}}_k$ denotes the gene expression effect of $\mathbf{g}^{\overline{M}}_k$on clinical outcome Y, called a type $\overline{M}$ effect.

2.2.1.3 Full Bayesian model with missingness Figure 1b gives a full representation of our model. Our model contains three parts: *mechanistic model*, *clinical model*, and an *imputation model* derived from the previous two models. The mechanistic model and clinical model stem from the iBAG model introduced in Chapter 2.1.2. Here, we will focus on two novel parts added to the original iBAG model: the sparsity-induced spike-and-slab priors (Ishwaran and Rao, 2005) for feature selection in both mechanistic and clinical models, and an imputation model to deal with missing omics data.

There are a total of 8 groups of parameters that need to be estimated: γ^M, $\gamma^{\overline{M}}$, γ^C, Ω, σ^2_k's, $\mathbf{G}^{mis}$, $\mathbf{M}^{mis}$, σ^2. γ^M and $\gamma^{\overline{M}}$ are our parameters of interests. Investigators may also find Ωs important if they are interested in further inference on methylation-gene regulation. The corresponding Monte Carlo Markov Chain (MCMC) Gibbs sampler will be further discussed in Appendix.

The original iBAG model placed a Laplace prior on γ^M and $\gamma^{\overline{M}}$ for shrinkage purpose, however, it does not set the effects to exact zeros. To conduct natural variable selection, we instead use a spike-and-slab prior to induce sparsity. In addition, we also perform feature selection in ω using the same spike-and-slab prior, considering that not all the methylation sites are regulating the gene expression. So we will have:

$$\left[\gamma_k^M | I_k^M, (\tau^M)^2\right] \sim (1 - I_k^M)\delta_0(\cdot) + I_k^M N(0, (\tau^M)^2);$$

$$\left[I_k^M | \pi^M\right] \sim Bern(\pi^M), \ \left[\pi^M\right] \sim Unif(0,1);$$
$$\left[\gamma_k^{\overline{M}} | I_k^{\overline{M}}, (\tau^{\overline{M}})^2\right] \sim (1 - I_k^{\overline{M}})\delta_0(\cdot) + I_k^{\overline{M}} N(0, (\tau^{\overline{M}})^2);$$
$$\left[I_k^{\overline{M}} | \pi^{\overline{M}}\right] \sim Bern(\pi^{\overline{M}}), \ \left[\pi^{\overline{M}}\right] \sim Unif(0,1);$$
$$\left[\omega_j | I_j^\omega, (\tau^\omega)^2\right] \sim (1 - I_j^\omega)\delta_0(\cdot) + I_j^\omega N(0, (\tau^\omega)^2);$$
$$\left[I_j^\omega | \pi^\omega\right] \sim Bern(\pi^\omega), \ \left[\pi^\omega\right] \sim Unif(0,1);$$

where I_k^M, $I_k^{\overline{M}}$ and I_j^ω are binary indicators and $\delta_0(\cdot) = N(0, 10^{-6})$ represents a narrow *spike* and τ^2 represent the wide *slab*. A Jeffery's prior is put on $(\tau^M)^2$, $(\tau^{\overline{M}})^2$, $(\tau^\omega)^2$, i.e.:

$$\left[(\tau^M)^2\right] \propto (\tau^M)^{-2}, \left[(\tau^{\overline{M}})^2\right] \propto (\tau^{\overline{M}})^{-2}, \left[(\tau^\omega)^2\right] \propto (\tau^\omega)^{-2}.$$

For $\mathbf{G}_{mis}$ and $\mathbf{M}_{mis}$ (Figure 1b), we impose the following imputation model:

$$\mathbf{g}_{mis,k} \sim MVN_{N_{mis}^G \times N_{mis}^G}(\mathbf{M}_{\mathcal{J}_k}\omega_k, \sigma_k^2 \mathbf{I}_{N_{mis}^G \times N_{mis}^G});$$

$$\mathbf{m}_{mis,j} \sim MVN_{N_{mis}^M \times N_{mis}^M}(0, (\sigma^m)^2 \mathbf{I}_{N_{mis}^M \times N_{mis}^M}),$$

where MVN denotes the multivariate normal distribution and $(\sigma^m)^2 = 1$ if the methylation value is already standardized with mean 0 and standard deviation 1. Note that here we assume the methylation data are in M *value* according to Pan et al., 2010. If β *value* is to be used, we may need to replace the above prior with a truncated normal distribution bounded between 0 and 1. All other variables are given non-informative priors. Details can be found in the Appendix.

Remarks:

- In our model, we assume that the methylations are many-to-one mapped to genes, i.e., only the methylation within the promoter region of the gene is mapped to the gene, and each methylation will only be mapped to one gene. The mapping is also assumed to be known from biology background knowledge in our model. The methylation level is centered around 0 if we are using m-value.
- It is reasonable and necessary to assume that if for all the ω_j where j are within the promoter region of gene k, we let $I_j^{\omega} = 0$, then we automatically have $I_k^M = 0$. I.e., $I_k^M \neq 0$ only when at least one methylation is selected.
- In cases where there are more than a few methylation sites mapped to each gene, principal component analysis (PCA) is recommended on methylation sites per gene, and the first few PCs explaining large proportion of variation can be used in place of original methylation level. This takes advantage of the high correlation between methylation sites.

2.2.2 Inference and evaluation

The full Bayesian hierarchical model in the last section allows fast Gibbs sampler. Full conditional formula for iterative sampling are shown in Appendix. The final parameter estimates are calculated by averaging stabilized MCMC iterations (i.e., removing the first B_r burn-in period in MCMC iterations). The burn-in period B_r is determined using Geweke's convergence diagnostics (Geweke et al., 1992). Geweke diagnostics aims to test whether the first $a\%$ and last $b\%$ of the MCMC iterations have equal mean, and thus decide whether the samples are drawn from a stationary distribution. As suggested by Geweke, the first 10% of MCMC total iterations are taken as burn-ins once the MCMC chain pass the diagnosis.

2.2.2.1 Prediction and feature selection The Bayesian integrative model generates two major inference outcomes: prediction and feature selection. For prediction, denote $\hat{\gamma}_{(b)}^C$, $\hat{\gamma}_{(b)}^M$, $\hat{\gamma}_{(b)}^{\overline{M}}$, and $\hat{\Omega}_{(b)}$ as the simulated parameter estimates from the b-th iteration. For a new sample with omics data ($\mathbf{C}_{new}$, $\mathbf{G}_{new}$, $\mathbf{M}_{new}$), we average prediction of y from the $(B - B_r)$ stable MCMCs by $\hat{y}_{new} = (\sum_{b=B_r+1}^{B} \hat{y}_{new}^{(b)})/(B - B_r)$, where $\hat{y}_{new}^{(b)} = \mathbf{C}_{new} \cdot \hat{\gamma}_{(b)}^C + \mathbf{M}_{new} \cdot \hat{\Omega}_{(b)} \cdot$

$\hat{\gamma}^{M}_{(b)} + (\mathbf{G}_{new} - \mathbf{M}_{new} \cdot \hat{\Omega}_{(b)}) \cdot \hat{\gamma}^{\overline{M}}_{(b)}$.

Next, we summarize feature selection indicators $I^{M}_{k,(b)}$ and $I^{\overline{M}}_{k,(b)}$ for gene k and the b-th MCMC to determine the set of genes predictive to outcome y. Given the b-th iteration, we define the selection indicator for gene k as $I_{k,(b)}$ so that gene k is selected if either the impact on outcome is through methylation ($I^{M}_{k,(b)}$) or not ($I^{\overline{M}}_{k,(b)}$). In other words, $I_{k,(b)} = (I^{M}_{k,(b)})\ OR\ (I^{\overline{M}}_{k,(b)} = 1 - (1 - I^{M}_{k,(b)})(1 - I^{\overline{M}}_{k,(b)})$ for $B_r + 1 \leq b \leq B$ and $1 \leq k \leq K$. To control FDR at gene level, we apply Bayesian FDR (BFDR) proposed by Newton et al., 2004:

$$BFDR(t) = \frac{\sum_{k=1}^{K} \hat{P}_k(H_0|D) d_k(t)}{\sum_{k=1}^{K} d_k(t)}$$

where $\hat{P}_k(H_0|D) = 1 - (\sum_{b=B_r+1}^{B} I_{k,(b)})/(B - B_r)$ is the posterior probability of gene k belonging to null hypothesis H_0 (i.e. gene k is not selected given observed data $\mathcal{D}$), $d_k(t) = I(\hat{P}_k(H_0|D) < t)$, and t is a tuning threshold. Given the definition of BFDR, the q-value of gene k can be defined as $q_k = min_{t \geq \hat{P}_k(H_0|D)} BFDR(t)$. This q-value will later be used for feature selection decision, which is comparable to frequentist approaches. Among selected genes, one may perform post hoc analysis to further investigate $c^{M}_{k} = (\sum_{b=B_r+1}^{B} I^{M}_{k,(b)})/(B - B_r)$ and $c^{\overline{M}}_{k} = (\sum_{b=B_r+1}^{B} I^{\overline{M}}_{k,(b)})/(B - B_r)$ and determine whether the impact of gene k to outcome is through methylation, non-methylation or both.

2.2.2.2 Benchmarks for evaluation

To evaluate the performance, the basic approach we compare to is the complete case (CC) analysis. For full Bayesian model with missingness, we will also choose to impute gene expression only (FBM$_G$), methylation only (FBM$_M$) or both (FBM$_{GM}$) when applicable. In simulation studies, we also perform analysis of the complete data (full) to examine the reduction of accuracy caused by missingness.

To benchmark performance of outcome prediction, we consider RMSE:

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y_i})^2 / N}$$

In simulation studies, after parameter estimates are obtained we generate a new testing dataset with large sample size ($N = 2000$) and compute RMSE. In real data analysis, we perform 50-fold cross-validation on complete cases for RMSE evaluation. In each iteration,

2% of complete-case samples are set aside as test dataset, all remaining data are used to perform CC and FBMs analyses. The parameter estimates are then applied to the test data for outcome prediction. After ten iterations, cross-validated RMSE can be evaluated on all complete-case samples.

To evaluate feature selection performance, we order genes by q-value, plot receiver operating characteristic (ROC) curves and calculate area under curve (AUC) in simulations since the underlying true predictive genes are known. For real data, since we do not know the true features, we treat the gene selection result from full data analysis (full) as a surrogate of gold standard and compare gene selection from CC (or FBMs) to the full data analysis by tracing the top number of selected genes on the x-axis (e.g x=100 top selected genes by CC and full) and the overlapped number from CC and full on the y-axis (Figure 4b). Comparing the curves of CC and FBMs, a higher curve closer to the diagonal line shows more similar gene selection to "full" and thus an indication of better performance.

2.3 SIMULATION STUDIES

2.3.1 Simulation schemes

To evaluate performance of our full Bayesian model with missingness, we perform simulation based on data structure described in Chapter 2. Specifically, the clinical and methylation data matrices are simulated from $N(0, 10)$ and $N(0, 1)$ with $N = (50, 100, 200, 500)$, $L = 2$ and $J = 2,000$. Each methylation site is randomly assigned to a gene, with the constraint that each gene contains at least one methylation site. In the mechanistic model, $I_j^{\omega} = 1$, $\omega_{jk} = 5$ and $\sigma_k^2 = 4$ for $1 \leq j \leq J$ and $1 \leq k \leq K$, where the total number of genes $K = 1000$. We then simulate gene expression matrices from $N(\mathbf{M}\Omega, diag(\sigma_1^2, ...\sigma_K^2))$. In the clinical model, 10 genes are randomly selected to impact clinical outcome through methylation and 10 randomly impact not through methylation (i.e. The I^M vector has 10 out of K genes equal one and the remaining are zero, and similarly for $I^{\bar{M}}$. The selected genes in I^M and $I^{\overline{M}}$ can possibly overlap). For selected genes in I^M and $I^{\overline{M}}$, the corresponding γ_k^M and $\gamma_k^{\overline{M}}$

are set to 10. The coefficients for clinical data γ_l^C $(1 \leq l \leq L)$ are also set at 10 and $\sigma^2 = 9$ to simulate clinical outcome Y.

After full multi-omics datasets are simulated, data with missingness are generated with $\alpha\%$ of samples with missing gene expression data and another non-overlapping $\beta\%$ of samples with missing methylation data. We simulate three scenarios of missingness: (I) Missing only gene expression data with (α, β)=(0.1,0), (0.2,0) and (0.5,0); (II) Missing only methylation data with (α, β)=(0,0.1), (0,0.2), (0,0.5); (III) Non-overlapping samples missing either gene expression or methylation data with (α, β)=(0.1,0.1), (0.2,0.2), (0.3,0.3). For Scenario I, we evaluate CC and FBM$_G$ approaches and compare with full. Similarly for Scenario II, we compare CC, FBM$_M$ and full. Finally for Scenario III, we compare CC, FBM$_G$, FBM$_M$, FBM$_{GM}$ and full. In this case, FBM$_G$ imputes gene expression but ignores samples with missing methylation and similarly, FBM$_M$ imputes methylation but ignores samples with missing gene expression. FBM$_{GM}$ utilizes all samples and imputes both gene expression and methylation.

2.3.2 Results

Figure 2 shows the outcome prediction performance by RMSE for all three scenarios. We first focus on small sample size situations $N = 50 - 200$. In Scenario I, CC, FBM$_G$ and full have similar performance when $\alpha = 10\%$ missing but FBM$_G$ clearly outperforms CC when missingness increases to 20% and 50%, showing the benefit of imputation as expected. In contrast, FBM$_M$ performs much worse than CC in Scenario II with $\beta = 10\%$ methylation missingness and FBM$_M$ only slightly outperform CC when missingness increases to 50%. Results of Scenario III are consistent with results of Scenarios I and II. FBM$_M$ and FBM$_{GM}$ performs worse than CC at $\alpha = \beta = 10\%$. FBM$_M$, FBM$_G$ and FBM$_{GM}$ outperforms CC at $\alpha = \beta = 50\%$. It is worth noting that when sample size increases to $N = 500$, the data information is strong enough such that CC has performance similar to full. Imputation almost always create more data uncertainty and have worse performance than CC, especially since the majority of the imputed data are irrelevant to the clinical outcome.

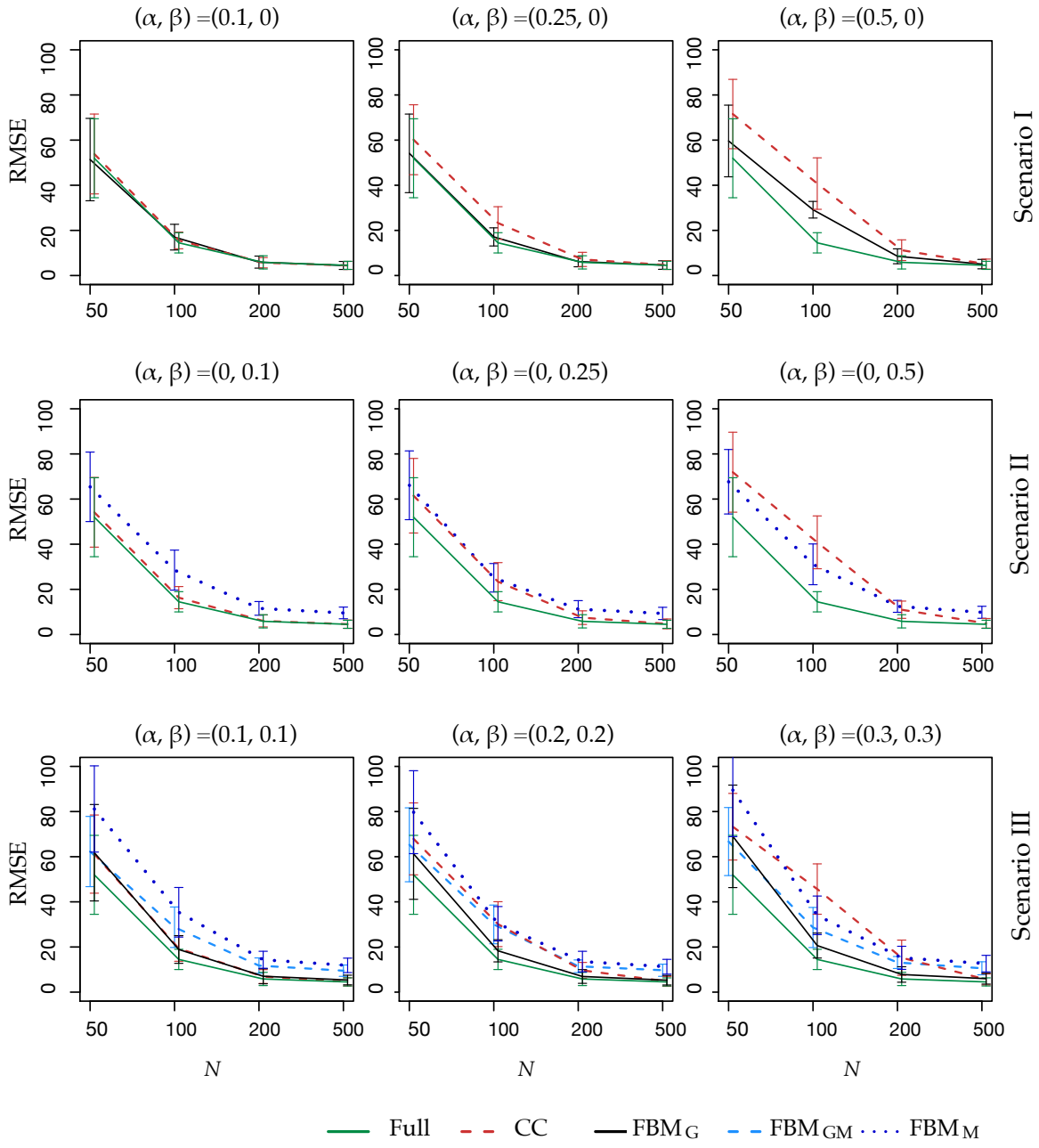


Figure 2: Model prediction by $RMSE(\hat{y})$ comparing different methods and full data.

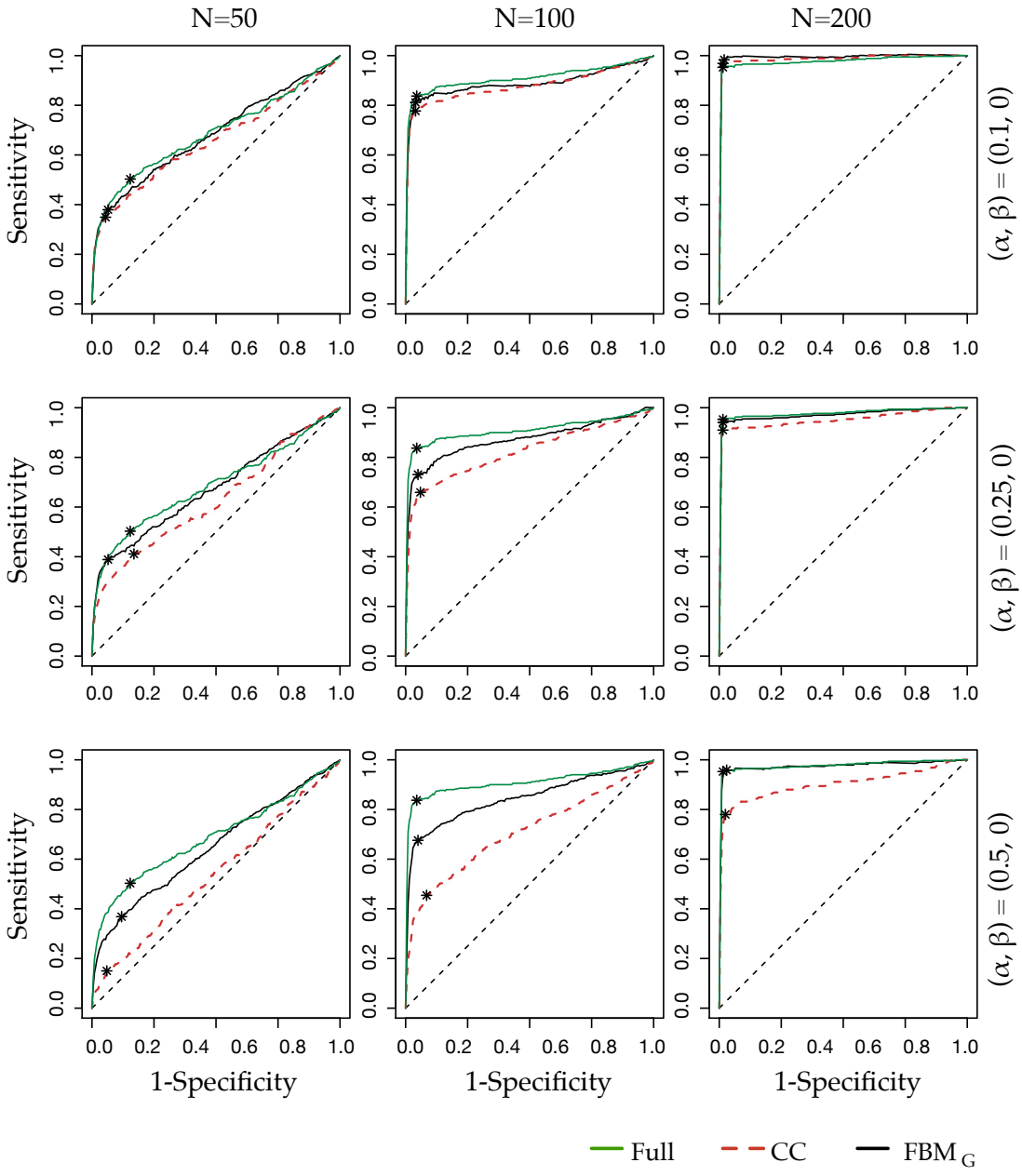


Figure 3: The ROC curves for feature selection comparison on Scenario I. $\alpha \neq 0, \beta = 0$. Star on each ROC curve is the point with maximum Youden index.

We next examine feature selection performance by AUC values in Table 1 (ROC curves shown in Supplement Figure 20a-c). In Scenario I, FBM_G always has higher AUC values than CC especially when missingness α increases to 25% or 30%. On the contrary, FBM_M has lower AUC than CC except for $N = 50, 100$ in $\beta = 50\%$. The message becomes mixed in Scenario III as expected.

In summary, since methylation is the up-regulator of gene expression and indirectly impact the clinical outcome, imputing methylation is generally less effective than imputing gene expression. With increasing missing proportion, the benefit of imputation escalates. But when the missing proportion is small or sample size large, imputation may only introduce data uncertainty and the performance becomes worse than complete case.

2.4 IMPUTE OR NOT: A DECISION SCHEME BY CROSS-VALIDATION

From the mechanistic model in Figure 1b, gene expression and methylation data are not symmetric. Methylation data can be predictive to gene expression data and further predictive to outcomes. On the other hand, gene expression data are less predictive to methylation to help outcome prediction. As we have shown in simulations, imputations do not always improve prediction performance compared to complete case analysis. Whether imputation would improve outcome prediction depends on the types of missingness (missing the upstream methylation data or missing the downstream gene expression data or both), the proportion of missingness and sample size. To guide the decision, we propose a self-learning cross-validation (CV) scheme. Specifically, we apply 10-fold cross-validation by leaving 10% of complete-case samples as the test set (for Scenario III N=500, we apply 50-fold CV). We apply CC, FBM_G, FBM_M and FBM_{GM} to the remaining training data, calculate parameter estimates and apply to test set. The procedure is repeated across all 10 folds and RMSE can be calculated on all test sets. The method with the smallest RMSE is selected to determine whether and how to impute. We note that to evaluate performance of CV scheme in RMSE evaluation of outcome prediction in the Chapter 5, nested cross-validation will be used (i.e. An outer loop of cross-validation is used to evaluate RMSE and an inner loop of CV scheme

Table 1: AUC of different methods in simulation studies

			Scenario I		Scenario II		Scenario III			
Missing	N	Full	CC	FBM_G	CC	FBM_M	CC	FBM_G	FBM_{GM}	FBM_M
Low*	50	0.676	0.649	0.672	0.631	0.613	0.616	0.628	0.652	0.583
	100	0.899	0.867	0.878	0.883	0.777	0.826	0.753	0.826	0.703
	200	0.967	0.976	0.983	0.961	0.863	0.946	0.859	0.913	0.789
Med*	50	0.676	0.619	0.665	0.626	0.597	0.574	0.61	0.646	0.528
	100	0.899	0.813	0.862	0.807	0.774	0.762	0.745	0.826	0.699
	200	0.967	0.941	0.961	0.939	0.861	0.903	0.858	0.917	0.795
High*	50	0.676	0.53	0.637	0.557	0.588	0.534	0.583	0.636	0.512
	100	0.899	0.699	0.835	0.681	0.742	0.638	0.75	0.833	0.7
	200	0.967	0.892	0.964	0.88	0.814	0.828	0.782	0.883	0.765

*Low, medium, and high missing proportion for Scenario I and II are 10%, 25% and 50% for either α or β, respectively; and for Scenario III is 10%, 20% and 30% for both α and β.

An outer loop of cross-validation is used to evaluate RMSE and an inner loop of CV scheme for method selection is performed in each training set of the outer loop).

Figure 4 shows scatter plot of RMSE performance comparing FBM_G and CC in Scenario I (Figure 4a; $\alpha = 50\%$ missing gene expression) and FBM_M and CC in Scenario II (Figure 4b; $\beta = 50\%$ missing methylation) in 50 independent simulations. In Scenario I,FBM_G generally has smaller RMSE than CC in small sample sizes ($N = 50$, 100, 200). But for $N = 500$, RMSE of FBM_G becomes slightly larger than CC on average. For missing methylation in Scenario II, FBM_M performs better than CC at $N = 50$ but gradually becomes worse than CC at $N = 100$, 200 and 500. We applied the CV scheme in each simulation to determine whether imputation should be performed or not. Simulations shown by circles represent correct decisions (i.e. CV scheme decides to impute and the RMSE of imputation is indeed smaller than RMSE of CC or vice versa) and cross represents incorrect decision. The result shows universally high accuracy of CV scheme decision. When the decision is wrong, imputation and CC RMSEs are close to each other (near the diagonal line) and the incorrect decision only minimally impacts the outcome prediction. Figure 4c shows box plots of RMSE generated from different approaches (CC, FBM_G, FBM_M , FBM_{GM} , CV) for Scenario III ($\alpha = \beta = 25\%$). At $N = 50$, FBM_G and FBM_{GM} both perform well and CV generates similar small RMSE. When N increases to 500, FBM_{GM} becomes much worse and CC performs slightly better than FBM_G. Again, the CV scheme makes mostly correct decision and thus generates small RMSE close to the lowest. In conclusion, the simulation results indicate effectiveness of the CV scheme in determining the best strategy of whether and how to impute when encountering missingness in multi-omics data.

2.5 REAL APPLICATION

2.5.1 Data and approach

We apply the proposed full Bayesian model with missingness to 460 children asthma nasal epithelium samples obtained from asthma study at Children's Hospital of Pittsburgh, with complete DNA methylation data from Illumina 450k chips and RNA-Seq gene expression

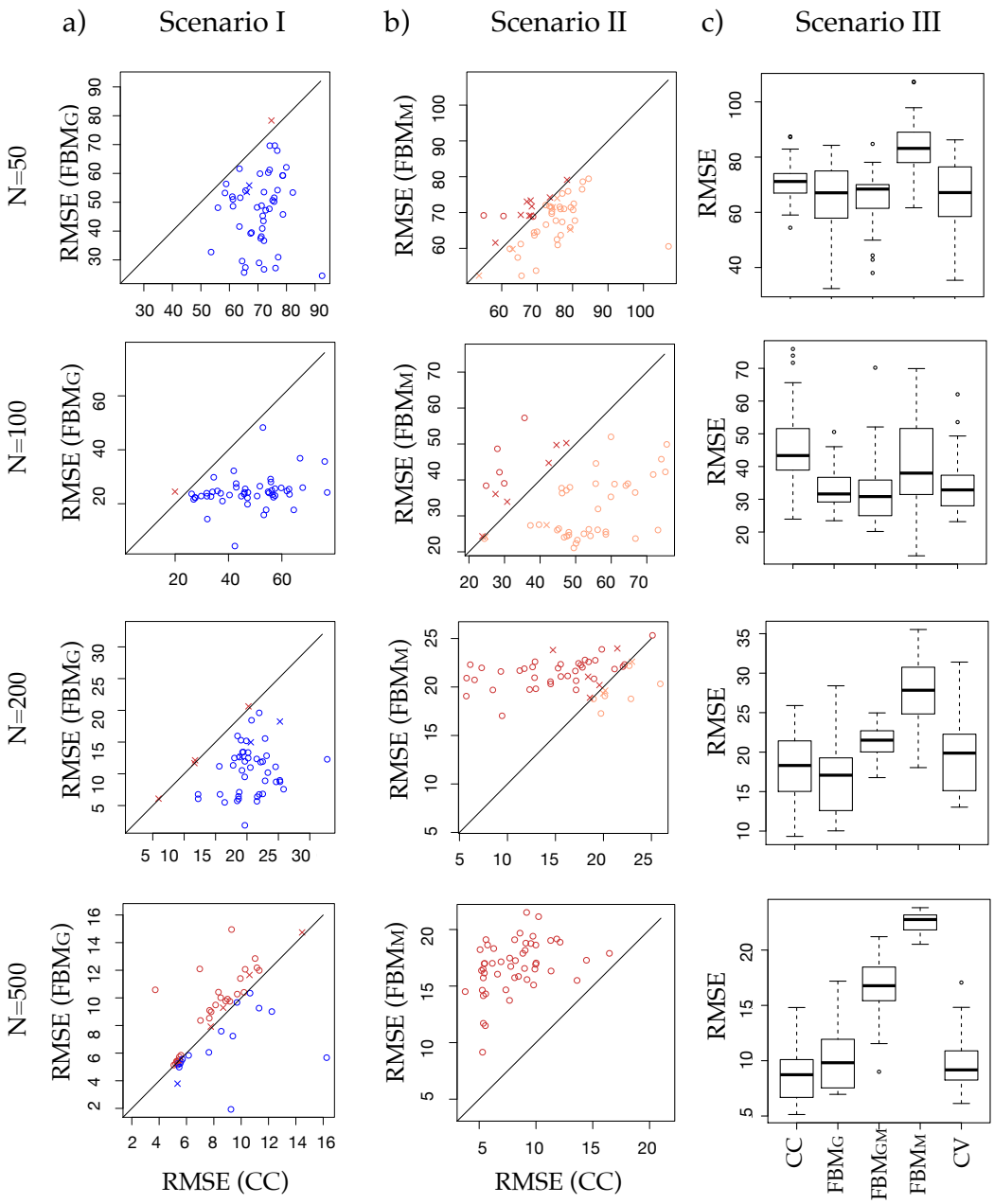


Figure 4: Model prediction performance of CV scheme.

a-b) In missing Scenario I ($\alpha = 50\%, \beta = 0$) and Scenario II ($\alpha = 0, \beta = 50\%$), scatter plot of RMSEs between two methods are shown, where circle means cross-validation scheme generates correct decision, and cross means mistakes. Sample size N varies from $50, 100, 200, 500$. c) Box plot of RMSE of different methods and CV selection scheme in Scenario III ($\alpha = \beta = 25\%$).

data. All data were preprocessed with standard procedures and bioinformatics tools. We used M-value for methylation level for better model fitting. The RNA-seq gene expression counts were transformed to TPM (transcripts per million), a continuous value also more valid for the model assumptions. We filter out genes with small mean expression level (< 130.41) or small standard deviation (< 23.83) to obtain $K = 1,000$ genes for the analysis. We then select methylation sites matched to these 1,000 genes. Since some genes have many corresponding methylation sites, we perform principal component analysis (PCA) to identify the top eigen-methylation sites as input to the model. We take no more than 3 PCs per gene or less than 3 PCs that explain at least 75% variation. This generates $J = 2619$ eigen-methylation features for the analysis. The PCA analysis reduces redundant (highly correlated) information in methylation sites and independence of eigen-methylation features fits the model assumptions well. Similar to simulation, we generate three scenarios of missingness: (I) $\alpha = 50\%$, (II) $\beta = 50\%$, (III) $\alpha = \beta = 20\%$. For each scenario, we repeat 50 times.

Serum Immunoglobulin E (IgE) level is a primary clinical outcome in children asthma studies. We take log-transformed IgE level as our clinical outcome, and age and gender as clinical variables. Together with the gene expression and methylation PCs, we run our model with full data and three scenarios of missingness to compare complete case approach (CC) and full Bayesian model with missingness (FBM_G, FBM_M and FBM_{GM}).

2.5.2 Outcome prediction and feature selection

Table 2 shows the RMSE of outcome prediction from complete case analysis and FBMs. When 50% samples have missing gene expression (Scenario I), FBM_G had reduced RMSE compared to CC (dropped from 43.34.19 to 36.04). In contrast, FBM_M had inflated RMSE compared to CC when 50% of missing methylation (51.59 compared to 39.54 in Scenario II). When gene expression and methylation both missed 20% of samples (i.e. 40% of samples had missing values in Scenario III), FBM_G had the smallest RMSE (38.71) compared to FBM_{GM} (46.15), FBM_M (54.23) and CC (39.09). Our CV scheme performed the automatic selection of whether and how to impute, and the RMSE was always close to the best (I: 36.61, II: 41.09

Table 2: RMSE (S.E.) of different methods

Methods	Scenario I	Scenario II	Scenario III
CC	43.34 (5.33)	39.54 (6.85)	39.09 (5.43)
FBM_G	36.04 (4.73)		38.71 (5.73)
FBM_{GM}			46.15 (4.74)
FBM_M		51.59 (6.71)	54.23 (4.50)
CV	36.61 (3.78)	41.09 (6.17)	40.62 (7.36)

and III: 40.02). Figure 5a shows scatter plot or box plot of RMSEs of different methods in all three Scenarios. Similar to the simulation result, CV scheme mostly selected the best method and the mistakes were near the diagonal line with little predictive impact.

Unlike in simulation, no underlying truth is available for real data and thus calculation of AUC is not possible. Figure 5b treats the predicted outcome from full data as the surrogate of underlying truth and compare feature selection from each method with the surrogate (i.e. x-axis shows the same number of features selected by the designated method and full and y-axis demonstrates the overlap between the two). A curve with higher overlap shows better similarity of feature selection with the surrogate, an indication of better performance. Similar to the RMSE result, FBM_G performed better than CC in Scenario I, CC better than FBM_M in Scenario II, and FBM_G and CC performed better than FBM_M and FBM_{GM}in Scenario III. CV scheme performs close to the best. All results from this real application are largely aligned with our observations in simulation studies.

We next investigate whether feature selection from the imputation methods or method selected by CV scheme represent better functional annotation in a biological sense. We obtained the top 200 genes from feature selection of each method by posterior probability order and then performed pathway enrichment analysis by one-sided Fisher's exact test. We collected 2,467 pathways from four pathway databases (KEGG, Reactome, Biocarta and

GO) with the restriction of pathway size between 10 to 500. The enrichment p-values were then adjusted for multiple comparisons by Benjamini-Hochberg procedure. We found 27 enriched pathways from the full data analysis under FDR=1%. We used these 27 pathways as a surrogate of gold standard to benchmark functional annotation performance of each method. Figure 6 shows box-plots of the minus log-transformed p-values from pathway enrichment of the 27 pathways in each method. As expected, FBM_G had better p-value significance distribution than CC in Scenario I. In Scenario II, CC performed better than FBM_M. FBM_{GM} and FBM_G outperformed CC in Scenario III. The CV scheme automatically determined whether and how to impute, and it always perform close to the best in each scenario.

2.6 DISCUSSION

Integrative analysis of multi-level omics data brings unique insights to the modulating relationship between different types of omics data. Feature selection and model prediction are two important goals in multi-omics integration, which empowers discovery of disease-associated biomarkers, survival prediction, and risk assessment. Several methods have been developed to fulfill these goals, including iBAG using a two-layer Bayesian hierarchical model to discover both association between genes and clinical outcome, and that between gene and upstream regulators. However, none of these methods are able to handle the potential large proportion of missing data in the data integration. In this paper, we propose a full Bayesian model with missingness (FBM) inspired by iBAG model, to jointly perform feature selection, model prediction, and missing data incorporation. In addition to the mechanistic model and clinical model originally proposed by iBAG, FBM includes a third layer of missingness model to incorporate samples with missingness. The flexibility of Bayesian hierarchical modeling and Gibbs sampler technique enables us to jointly model association among data and infer parameters in all three layers of models together. The Laplace (double exponential) prior initially used in iBAG could not realize exact feature selection. FBM applied spike-and-slab prior for more effective feature selection and allows Bayesian FDR control. We demonstrated outcome prediction and feature selection performance of FBM using extensive simulations.

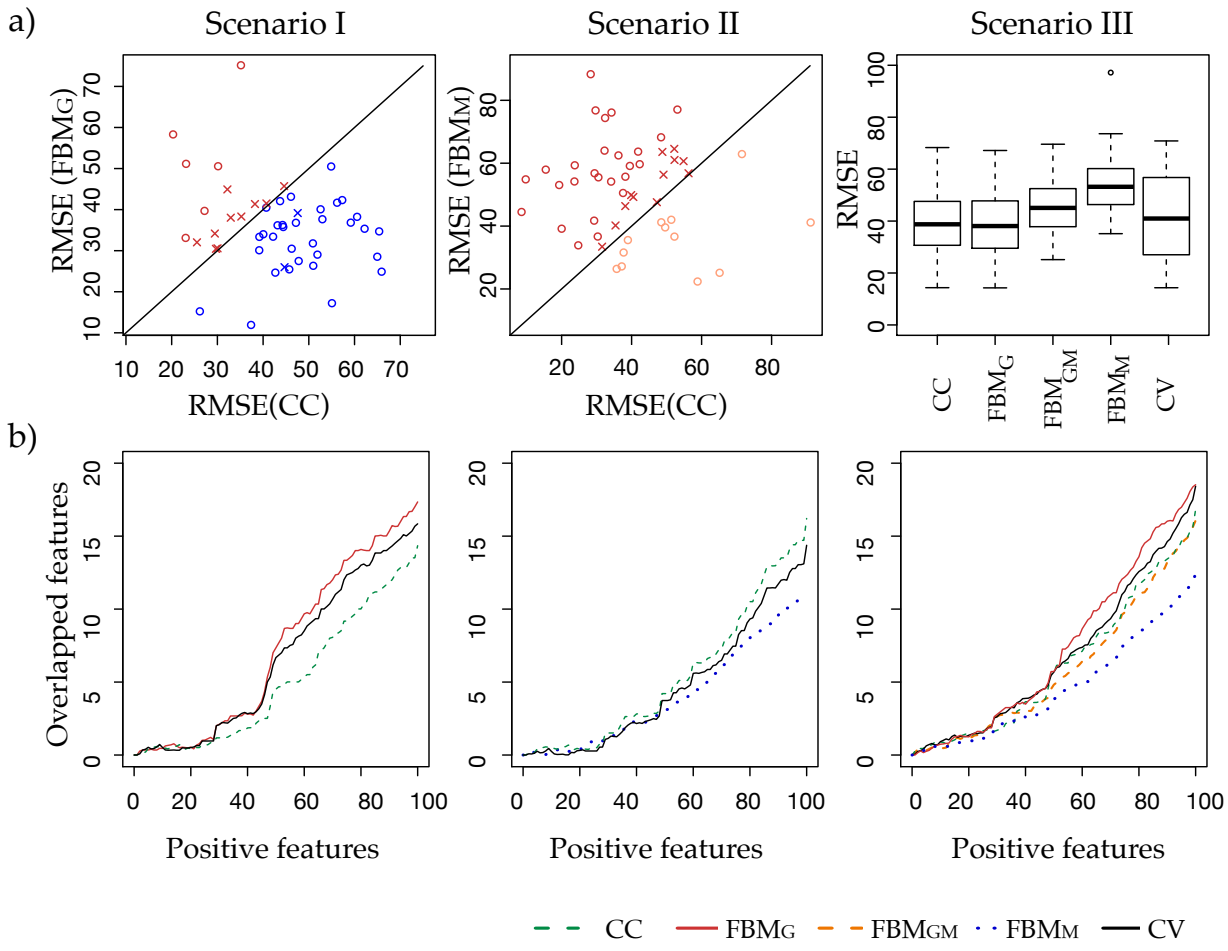


Figure 5: Model prediction and feature selection of CV scheme in real data.

a) scatter plot of RMSEs between two methods are shown, where circle means cross-validation scheme generates correct decision, and cross means mistakes. b) The number of overlapped genes (features) selected by each methods compared with full asthma data. On x axis is the top number of genes selected by each method.

From the extensive simulations, we then further realized that imputation is not always favored over complete case analysis. For example, in high dimensional genomic data analysis, when missingness exists in upstream regulators (i.e. methylation), uncertainty will be increased to impair final inference. When the missing proportion is relatively small and the

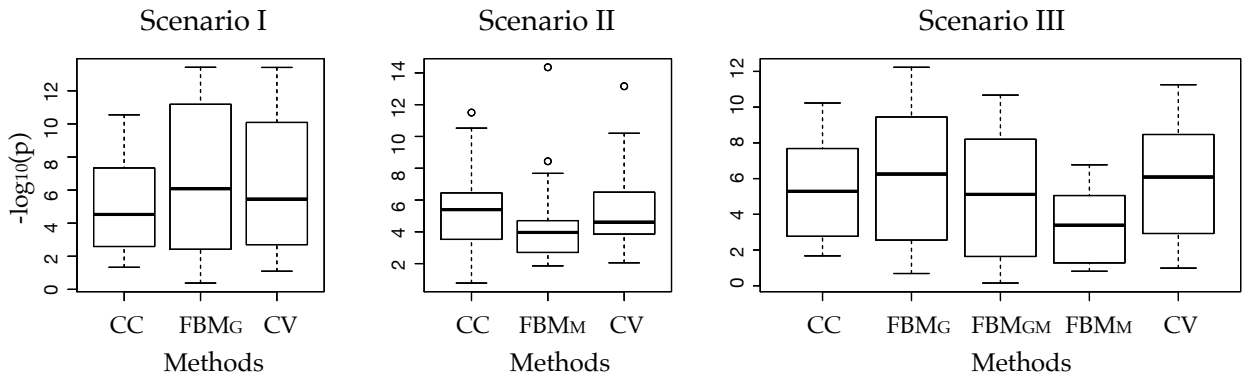


Figure 6: Pathway analysis similarities between methods with full data.

Compare different methods for the $-log_{10}(p)$ value of the top 27 pathways taken from full data with q-value <0.01.

sample size is large (i.e. signal is strong), complete case analysis also often outperforms FBM. To decide the best handling of data with missingness, we proposed a self-learning cross-validation decision scheme. Previously, we have developed a similar self-training selection scheme for selecting the best microarray missing value imputation method and its downstream biological impact (Brock et al., 2008; Oh et al., 2011). In both simulation and the childhood asthma application, we showed superior performance of the CV scheme in prediction outcome and feature selection.

While Bayesian hierarchical model allows complex parameter structures, it also comes with a computational cost when using conventional inference approaches such as Metropolis-Hasting or its special case, Gibbs sampling. In FBM, fast Gibbs sampling was applicable using conjugate priors and the convergence was generally fast (B=2000). We have optimized the R code using C++ with Rcpp package. The computing takes 90 minutes for a reasonably large dataset with $N = 500$ samples, $K = 1000$ genes, $J = 2000$ methylations and $B = 2000$ MCMC iterations using a regular computer with 1 Intel Xeon CPU (2.40 GHz). Since the computation burden grows linearly with sample size, feature size and number of iterations, applying FBM with parallel computing could be a solution to substantially speed up computing and allow routine omics applications. An R package, data and source code to replicate all results in this paper are available on GitHub (https://github.com/CHPGenetics/FBM).

3.0 INCORPORATING LINKAGE DISEQUILIBRIUM AND IDENTITY-BY-DESCENT INFORMATION TO IMPROVE GENOTYPE CALLING FROM FAMILY-BASED SEQUENCING DATA

3.1 INTRODUCTION

Next generation sequencing (NGS) technologies has greatly facilitated the discovery and detection of genetic variants, serving as fuels for further downstream analysis. For instance, the genome-wide associate studies (GWAS) have discovered thousands of associations between a variety of traits and diseases and single nucleotide polymorphisms (SNPs).

One key first step in NGS downstream analysis is the inference of genotypes given the raw sequencing reads or another preliminary inference. This inference step is ordinarily called genotype calling or genotyping. While genotype calls with sufficient accuracy for further downstream analysis could be achieved by a redundant sequencing on relatively limited sample size, or referred as deep sequencing, many have proven that for a given budget, sequencing more samples with shallower depth, facilitated by likelihood-based LD-aware methods, could greatly improve genotype calling accuracy (Li et al., 2011; Li and Stephens, 2003). For example, a deep sequencing with 30X read depth typically results in $> 99\%$ overall genotype calling accuracy. Yet Li et al. (2009) showed that, under similar budget, with the help of LD-aware genotyping methods, sequencing more individuals with read depth as low as 2X achieved higher overall genotyping accuracy compared with deep sequencing less individuals with 30X read depth (Li et al., 2011). This low-coverage large-cohorts sequencing strategy has been widely adopted by genotype profiling projects, including UK10K project and 1000 Genome Project. Correspondingly, genotyping methods designed for low-depth sequencing are emerging. These methods either using dynamic programming technique like

hidden Markov model (HMM) or EM algorithms, and are largely based on the LD pattern recognition (Li et al., 2011, 2010; Browning and Browning, 2016; Chen et al., 2013; Chang et al., 2016; Zhou and Whittemore, 2012).

On the other hand, family-based sequencing projects have their unique strengths compared to unrelated population sequencing in, for example, detecting rare causal variants (MAF $< 1\%$), understanding parent-of-origin effects, and studying Mendelian disorder. When sequencing from the family-based data, genotype calling could be further improved by leveraging partially the pedigree information, or identity-by-descent (IBD) information, in the statistical models (Li et al., 2015; Chen et al., 2013; Chang et al., 2016; Zhou and Whittemore, 2012). For example, Chen et al. (2013) has benchmarked that by considering trio parent-offspring constraints imposed by Mendelian inheritance, the genotype calling in trio family-sequencing data is improved, compared with previous LD-based methods ignoring the family constraints. Chang et al. (2016) extended the idea to general family structures by breaking down pedigree to trios and achieved higher genotyping accuracy in such family structures compared with its predecessor. And Zhou and Whittemore (2012) took both family constraints and LD constraints into account in a EM-algorithm based likelihood model, and showed the gain in genotyping accuracy by leveraging both LD and pedigree information. Yet this program is not publicly available, and compared with hidden Markov models, EM algorithm is not a mainstream solution for genotyping problems. Li et al. (2015) proposed a method that uses inheritance vectors (IV) to describe and IBD patterns, therefore improve genotype calling in family-based data for especially rare variants. Yet an effective, accurate, and computationally feasible method to comprehensively incorporate both LD and pedigree constraints for genotype calling is still lacking.

Here we propose our HMM-based model, inspired by aforementioned existing methods, to improve genotype calling in family-based sequencing data. In our model, we jointly infer the genotyping likelihood in a family with both LD patterns and IBD patterns considered. We benchmarked the performance of our method using both synthetic data (real data simulation using 1000 Genome Project) under different sequencing scenarios with varies of sequencing depth (2X-10X) and per-base error rate (Phred score 20 and 30). For comparison, we also benchmarked a few of the most popular or best performing methods, GATK, Bea-

gle4, FamLDCaller, Polymutt2. The results show that our method achieves highest overall genotype calling accuracy, and the improvement under some scenarios could be up to 50%. The main gain in genotyping accuracy happens in rare (MAF $< 1\%$) and moderately rare allele (MAF 1-5%). We also apply our method to a family-based whole exome sequencing dataset from anorexia cohort, collected in Kartini Clinic in Portland, Oregon, with 74 individuals within 26 pedigrees. We randomly subsample sequencing reads to mimic shallower sequencing scenarios to the average sequencing depth from 4X to 9X. The results showed that LDIV has the highest genotyping accuracy and relatively low Mendelian error rates.

3.2 METHODS

Similar to Li et al. (2015), we start introducing genotype inference by assuming the ith genome position is covered by N mapped A, C, G, T bases w/ counts N_A, N_C, N_G, and N_T. Base error rate e_i correspond to mapped base calls $b_i \in \{A, C, G, T\}$ indicate this location is incorrectly called. R_i to denote all the base calls, which are the observed read data, and G_i, the hypothetical true genotype. Due to sequencing errors and allele dropout, a total of $\binom{4}{2} + 4 = 10$ genotypes are possible.

We have two assumptions 1) all base calling errors are independent; and 2) for heterozygous genotypes, we assume each base originates from either one allele with equal probabilities. E.g. $P(A|AC) = P(C|AC) = 0.5$.

Base b_i is either A or C then $P(b_i|AC) = \frac{1}{2}(1 - 2e_i/3)$ and otherwise $P(b_i|AC) = e_i/3$. Then the GL (genotype likelihood) of genotyp AG combining all bases is

$$P(R_i|AC) = \prod_{k=1}^{N_A+N_C} \frac{1}{2}(1 - 2e_k/3) \prod_{k=1}^{N-N_A-N_C} (e_k/3)$$

Then we can calculate all possible GL by traversing all 10 genotypes.

3.2.1 Leveraging LD Information with Hidden Markov Model

Li et al. (2011) developed an LD-based computational method to increase genotyping accuracy with linkage disequilibrium information. Chen et al. (2013) and Chang et al. (2016) extended that methods to include partial information from the family structure. Those methods are essentially describing the entire sequenced haplotypes as if they are constructed by mosaics from multiple other reference haplotypes, either from other family members, or from outside reference panels. First we define $P(G_i|H_i)$ as the probability of an underlying true genotype G_i given mosaic chromosome H_i, the state in HMM. To calculate this we define $T(G_i)$ or $T(H_i)$, the number of variant alleles in G_i or in the reference haplotypes indexed by H_i. To be specific, genotype {0,0}, {0,1}, {1,0}, and {1,1} correspond to T = 0,1,1,2, respectively.

$$P(G_i|H_i) = \begin{cases} (1-\varepsilon_i)^2 & \{T(H_i)=0 \text{ or } T(H_i)=2\} \text{ and } T(H_i)=T(G_i), \\ \varepsilon_i(1-\varepsilon_i) & \{T(H_i)=0 \text{ or } T(H_i)=2\} \text{ and } |T(H_i)-T(G_i)|=1, \\ \varepsilon_i^2 & \{T(H_i)=0 \text{ or } T(H_i)=2\} \text{ and } |T(H_i)-T(G_i)|=2, \\ (1-\varepsilon_i)^2+\varepsilon_i^2 & T(H_i)=1 \text{ and } T(H_i)=T(G_i), \\ 2\varepsilon_i(1-\varepsilon_i) & T(H_i)=1 \text{ and } T(H_i)\neq T(G_i) \end{cases}$$

where ε_i is the mosaic error rate representing cumulative effects of mutation and gene conversion. Together $P(R_i|G_i)$ and $P(G_i|H_i)$ give us:

$$P(R_i|H_i) = \sum_{G_i} P(R_i|G_i)P(G_i|H_i) \tag{3.2.1}$$

This is because we have:

$$P(R_i|H_i) = \sum_{G_i} P(R_i, G_i|H_i) = \sum_{G_i} P(R_i|G_i, H_i)P(G_i|H_i)$$

$$P(R_i|G_i, H_i) = P(R_i|G_i)$$

The above equation stands because given genoypes alone we have the probabilities of reads, regardless of the reference states. I.e., reads and states are conditional independent given genotypes.

Or in notations in the *Thunder*, equation (3.2.1) is written as:

$$P(R_i|H_i) \propto P(R_i, H_i)$$

The above could be written out as the factorization of forward and backward variable in HMM.

Then we calculate the transition probability in HMM $P(H_{i+1}|H_i)$ as:

$$P(H_{i+1} = (w, v)|H_i = (x, y)) =$$

$$\begin{cases} \theta_i^2/H^2 & x \neq w \text{ and } y \neq v \\ (1-\theta_i)\theta_i/H + \theta_i^2/H^2 & \{x = w \text{ and } y \neq v\} \text{ or } \{x \neq w \text{ and } y = v\} \\ (1-\theta_i)^2 + 2(1-\theta_i)\theta_i/H + \theta_i^2/H^2 & x = w \text{ and } y = v \end{cases}$$

where (x, y) and (w, v) denote indexes for the reference haplotypes at location i and $i+1$; θ_i denotes the mosaic transition rate and H the number of reference haplotypes Then $P(H_i|\mathbf{R})$ can be calculated using forward and backward (or called left and right) probabilities following HMM conventional algorithms such as Baum's forward-backward sampling algorithm. And our final goal, $P(G_i|\mathbf{R})$, can be calculated as following:

$$P(G_i|\mathbf{R}) = \sum_{H_i} P(G_i|H_i)P(H_i|\mathbf{R}) \tag{3.2.2}$$

by looping through all reference mosaics H_i. Similar to equation (3.2.1), the equation (3.2.2) stands because we have:

$$P(G_i|\mathbf{R}) = \sum_{H_i} P(G_i, H_i|\mathbf{R}) = \sum_{H_i} P(G_i|H_i, \mathbf{R})P(H_i|\mathbf{R})$$

and that:

$$P(G_i|H_i, \mathbf{R}) = P(G_i|H_i)$$

assuming genotypes and reads are conditional independent given states.

3.2.2 Leveraging IBD Information by Inheritance Vector

Li et al. (2015) proposed Polymutt2 seeking to improve genotype calling accuracy from the perspective of using IBD information to fully account for information from family structure. Assuming we start with ordered f founders and n non-founders, with M variants. Defining binary inheritance vector (IV) at variant i : $I_i = (p_1, m_1, ...p_n, m_n)$ where p_j and m_j indicating the transmission of paternal and maternal allele. 0, 1 indicates grand-paternal and grand-maternal allele being transmitted. There are total of $N = 2^n$ possible IVs, because half of the original $N = 2^{2n}$ possible IVs are redundant considering that we may simply flip each founder's haplotype. Let $v_k, k = 1, ...N$ denotes the kth IV; $\mathbf{R}_i$ and $\mathbf{G}_i$ being all reads and genotype at variant i, R_{ij} and $G_{ij} = (A_{father}, A_{mother})$ being reads and genotype of person j at variant i.

The likelihood of reads across M variants is:

$$P(R) = \sum_{I_1} ... \sum_{I_M} P(I_1) \prod_{i=2}^{M} P(I_i|P_{i-1}) \prod_{i=1}^{M} P(\mathbf{R}_i|I_i)$$

The initiation part is uniform across all possible IVs; transition part is calculated based on recombination rate; and the emission part is calculated as:

$$P(\mathbf{R}_i|I_i) \quad = \quad \sum_{G_{founders}} \prod_{j=1}^{f+n} P(R_{ij}|G_{ij}I_i) \prod_{j=1}^{f} P(G_{ij}) \tag{3.2.3}$$

Becaus we know genotype of all family members as long as the IV and genotype of founders are given. The prior probability $P(G_{ij})$ can be acquired from external sources or estimated assuming HWE holds.

One key goal here is to infer posterior distribution of the IVs, i.e. $P(I_i|\mathbf{R})$. We use forward-backward procedure in HMM, and the posterior probability of v_k at variant i is:

$$P(I_i = v_k|\mathbf{R}) \quad = \quad \frac{\alpha_i(k)\beta_i(k)}{\sum_{k=1}^{N} \alpha_i(k)\beta_i(k)} \tag{3.2.4}$$

From the marginal distribution we can find the inheritance vector with maximum posterior probability at variant i, denoted as I_i^{marg}. But this only maximize likelihood marginally at variant i. So we use Viterbi algorithm to infer a inheritance vector that achieve maximum likelihhod globally for all variants, denoted as I_i^{best}.

Then for the family we have our posterior likelihood as:

$$P(G_{i1},...G_{i(f+n)}|I_i^{best},\mathbf{R}_i) = \frac{P(\mathbf{R}_i|G_{i1},...G_{i(f+n)},I_i^{best})P((G_{i1},...G_{i(f+n)}|I_i^{best}))}{P(\mathbf{R}_i|I_i^{best})} = \frac{\prod_{j=1}^{f+n}P(R_{ij}|G_{ij},I_i^{best})\prod_{G_i^{founder}}P(G_i^{founder})}{P(\mathbf{R}_i|I_i^{best})} \quad (3.2.5)$$

3.2.3 Jointly Leveraging LD and IBD: LDIV

Here we propose our new method to jointly using LD and IBD information, which is expected to give us a more accurate overall (especially on not-rare variants) genotype estimation. A very generic form of fusing the two information is to be able to infer genotype conditional on both, or either, hidden states H (reference haplotypes) and inheritance vectors I. First we consider an individual at variant i, be reminded that our final goal is to get: $P(G_i|\mathbf{R})$. We assume that inheritance vectors and reference states are independent from each other given reads, i.e. $P(H_i|I_i,\mathbf{R}) = P(H_i|\mathbf{R})$, though this is not rigorously the case.
First in a certain family, for variant i:

$$P(G_{i1},..G_{i(f+n)}|I_i^{best},H_{i1},..H_{if},\mathbf{R}_i) = \frac{\prod_{j=1}^{f+n}P(R_{ij}|G_{ij},I_i^{best})\prod_{j=1}^{f}P(G_{ij}|H_{ij})}{P(\mathbf{R}_i|I_i^{best},H_{i1},..H_{if})} \quad (3.2.6)$$

The above equation stands because 1) given inheritance vector I_i^{best} we are able to pass the states and genotypes to each individual; and 2) given genotypes, reads and states are conditionally independent, i.e.:

$P(\mathbf{R}_i|G_{i1},..G_{i(f+n)},I_i^{best},H_{i1},..H_{if}) = \prod_{j=1}^{f+n}P(R_{ij}|G_{ij},H_{ij},I_i^{best}) = \prod_{j}^{f+n}P(R_{ij}|G_{ij},I_i^{best})$

Similarly, we have: $P(G_{i1},..G_{i(f+n)}|I_i^{best},H_{i1},..H_{if}) = \prod_{j=1}^{f}P(G_{ij}|H_{ij})$

For the denominator, similar to (3.2.3) we have

$$P(\mathbf{R}_i|I_i^{best},H_{i1},..H_{if}) = \sum_{G_{founders}}\prod_{j=1}^{f+n}P(R_{ij}|G_{ij},I_i^{best})\prod_{j=1}^{f}P(G_{ij}|H_{ij}) \quad (3.2.7)$$

The likelihood (3.2.3) and (3.2.7) differs only in the prior, which is the way founders' genotypes are inferred. Rather than a plain $P(G_{ij})$, LDIV incorporates the LD information by using states to correct possible errors of founder's genotype calls, i.e. $P(G_{ij}|H_{ij})$.

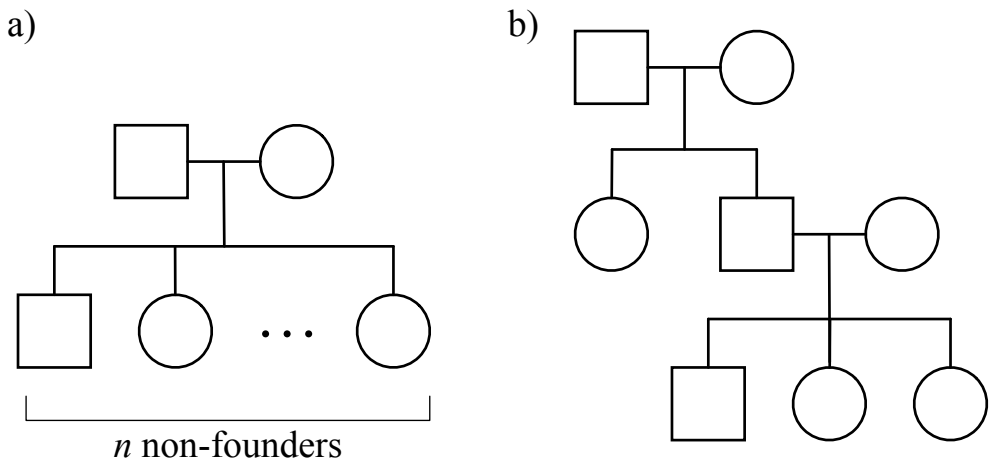


Figure 7: Family structures.

a) nuclear families, family type 1-4 corresponds to $n \in [1, 4]$; b) family type 5.

Plugging likelihood (3.2.7) into (3.2.6), we have:

$$P(G_{i1}, ..G_{i(f+n)}|I_i^{best}, H_{i1}, ..H_{if}, \mathbf{R}_i) = \frac{\prod_{j=1}^{f+n} P(R_{ij}|G_{ij}, I_i^{best}) \prod_{j=1}^{f} P(G_{ij}|H_{ij})}{\sum_{G_{founders}} \prod_{j=1}^{f+n} P(R_{ij}|G_{ij}, I_i^{best}) \prod_{j=1}^{f} P(G_{ij}|H_{ij})} \quad (3.2.8)$$

This is the final posterior likelihood. Note that the only differece between posterior likelihoods (3.2.5) and (3.2.8) lies in the prior, where Polymutt2 uses $P(G_{ij})$, LDIV will use $P(G_{ij}|H_{ij})$ instead. After we inferred I_i^{best} for $i = 1, ..M$, we take it as known. And for each states we will use above likelihood to update genotypes.

3.3 RESULTS

3.3.1 Simulation Scheme

To evaluate the performance of our method, we simulated genotypes from 150 families with 5 different family structures, including nuclear family with one up to four offspring, as well as 3-generation family with 5 non-founders (Figure 7). Each of our simulated data family

structure contains 30 families with one of the family structure. We generated approximately 10,000 sites from 1 million base pair length of haplotypes using 1000 Genomes Project data, mimicking the LD pattern and IBD pattern. Genotypes are sequenced at depth 2X, 6X and 10X to better represent the need in reality when only lower to middle depth sequenced data are provided for genotype calling. The per base error rate is set as 0.01 and 0.001, corresponding to Phred scaled based quality of Q20 and Q30, respectively.

Some of the methods including LDIV allow incorporating outside reference files to enhance genotyping performance by borrowing information from high confident calls. These reference files usually are usually in variant calling format (VCF), with high percentage of overlapping variants as the VCF files to be called, and are genotyped with high quality. Some real world reference files could be obtained from large sequencing projects like 1000 Genome projects. Here we adopt similar simulation scheme as aforementioned simulation data for simulating reference files, except that the source haplotypes for simulation data and reference data have no overlaps, mimicking the real world scenarios where individuals to be genotyped are not the same as the individuals from public genotyping references databases, although it is recommended that they are from same population.

We compare our method with other popular genotype inference tools including FamLDCaller, Polymutt2 and Beagle4 with recommended or default parameter settings on all simulated family-based sequencing data. To be specific, we set number of states at 40 with 30 rounds for FamLDCaller and our method LDIV; we set r2 parameter as 0.2 and minimum average depth as 2 for both Polymutt2 and LDIV; we set phasing iterations as 5 and imputation iteration as 5 for Beagle as suggested by their manuscript.

3.3.1.1 Benchmarking for Evaluation The most important benchmarking criteria is the genotyping accuracy. We compared all methods for genotype accuracy across all variants, as well as heterozygote sites only. And to gain a more in-depth view of the strengths and weakness of tools.to gaine also divided heterozygote sites markers by minor allele frequency (MAF) as representation of rare variants (MAF $< 1\%$), moderate variants (MAF 1-5%), and common variants (MAF $> 5\%$). Furthermore, we provided comparisons of another important criterions in family sequencing genotyping, the Mendelian error rates and phasing accuracies.

Taking the genotype from 1000 Genome project haplotypes as truth, the genotyping accuracy is defined as the number of correctly called loci divided by the total number of loci. And phasing accuracy is defined as the number of correctly phased loci divided by total number of correctly genotyped loci.

3.3.2 Simulation Results

3.3.2.1 Genotyping Accuracies Simulation results generally showed two trends (Figure 8-13, Supplementary Figure 23-25, Supplementary Table 15-19, 20-24). First is that for all methods, the overall genotyping accuracies increase when the sequencing depth increase (from 2X to 10X), as well as when the per-base error rate decreases (from 0.01 to 0.001). Another trend is that, when available, taking outside reference panels will generally increase genotyping accuracies. And this effect is most significant when sequencing depth is low, and per-base error rate is high, because the less information a sequenced individual can provide, the more boost in performance we expect to see when the same outside high quality information is borrowed.

The simulation results show the overall superior performance of LDIV compared to all other methods in overall genotype calling accuracy with or without reference files (Figure 8, 11, Supplementary Figure 23) for almost all family types. In very informative families like family type 4 (Figure 8) for some most contrasting scenarios for example, sequencing depth=6X, per-base error rate=0.01, LDIV without reference files has any overall genotyping accuracy of 99.93%, LDIV with reference files increases the accuracy to 99.97%; in contrast, the second best method in this case is FamLDCaller, which renders accuracies of 99.87% without reference file, and 99.92% with reference files. The largest reduce in genotype calling error rates are 46.2% without reference, and 62.5% with reference. When family structures are less informative, i.e., with less siblings and more complicated family structures which may hinder the construction of inheritance matrices (Figure 11), LDIV still achieves better overall genotyping accuracies in most of the cases, but the improvement is less than significant.

A further investigation of genotyping accuracy in only the heterozygote sites (Figure 9) show similar pattern as overall genotyping accuracy, LDIV is still the best performing

method, especially with reference panel included, which is as expected. For some of the most contrasting scenarios, the improvement in genotyping error rates of LDIV compared to the second best methods can be up to 60% with or without reference panels.

Methods using LD and IBD information is supposed to have different performance for rare and common SNPs, for example, methods leveraging LD information with the help of reference panel is supposed to be preferable for common SNPs, as those variants could be more likely found in the population where reference panel is sampled from. So we also look into the performance of different methods for rare (MAF $< 1\%$), moderate (MAF 1-5%), and common (MAF $> 5\%$) SNPs separately (Figure 10, 13). In almost all cases across all rarities, LDIV shows superior accuracy compared to other methods with or without the use of reference panel. This is because LDIV tend to weigh more on IBD information when family information is abundant (e.g. more siblings involved) and minor allele frequency is low; and tend to weigh on LD information when reference panel is provided.

3.3.2.2 Phasing Accuracies We then compare phasing accuracies of different methods, and we found that including reference files will usually, but not always increase the phasing accuracies (Table 4, Table 6, Supplementary Table 14). Although LDIV is not designed specifically for phasing purpose, its phasing accuracies are on a par with most of current methods. For example in family 5, with sequencing depth 2X, per-base error rate 0.01, LDIV will have phasing accuracy of 95.98% with reference files, and 97.03% without reference files, while the best method in this scenario is Beagle4, with 99.16%, and the worst is Polymutt2, with 86.12%. This is because Polymutt2 will rely solely on inheritance matrices for inference, following exact Mendelian law, and thus fail to take into account of similarity between individuals not related but might be from same population. Thus we can see that LD-aware methods is advantageous generally in phasing tasks compared to IBD-aware methods. Therefore, since LDIV takes reference from both IBD and LD information, its phasing performance is expected to be between two sets of methods.

3.3.2.3 Mendelian Error Rates Last but not least, we compare each methods performance in following Mendelian's Law by summarizing Mendelian error rates. Methods that

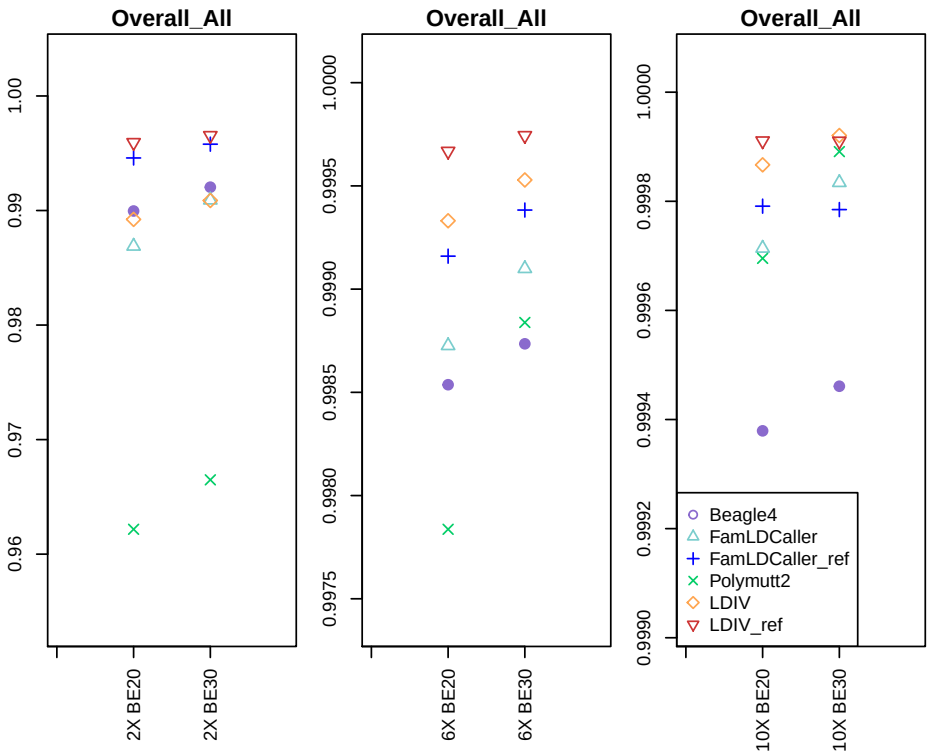


Figure 8: The overall genotyping accuracy for nuclear family type 4.

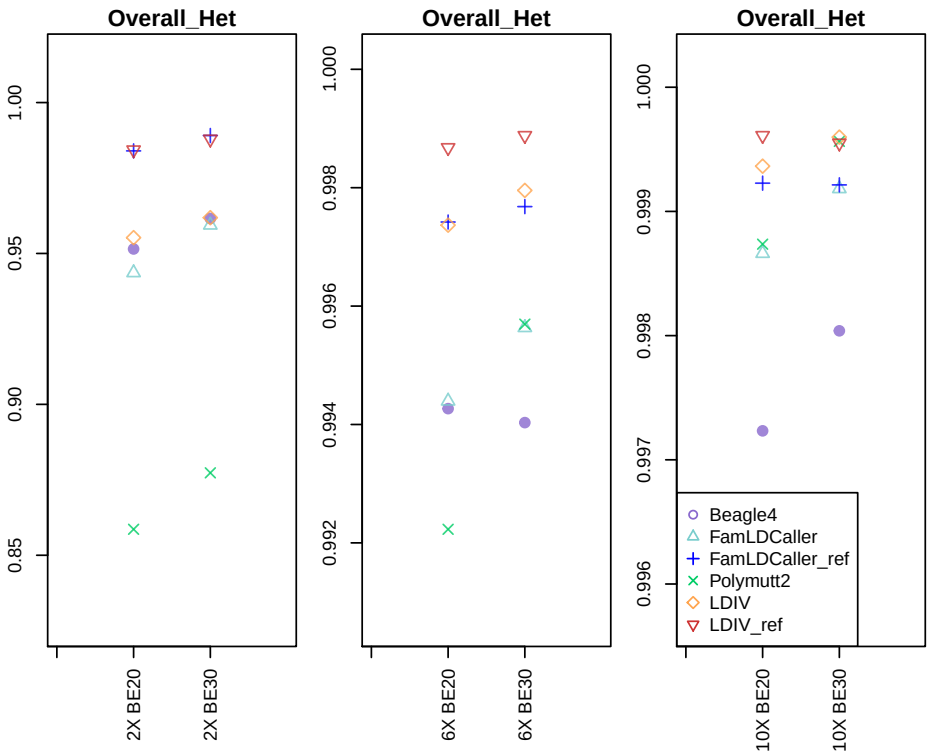


Figure 9: The genotyping accuracy on heterozygote sites for nuclear family type 4.

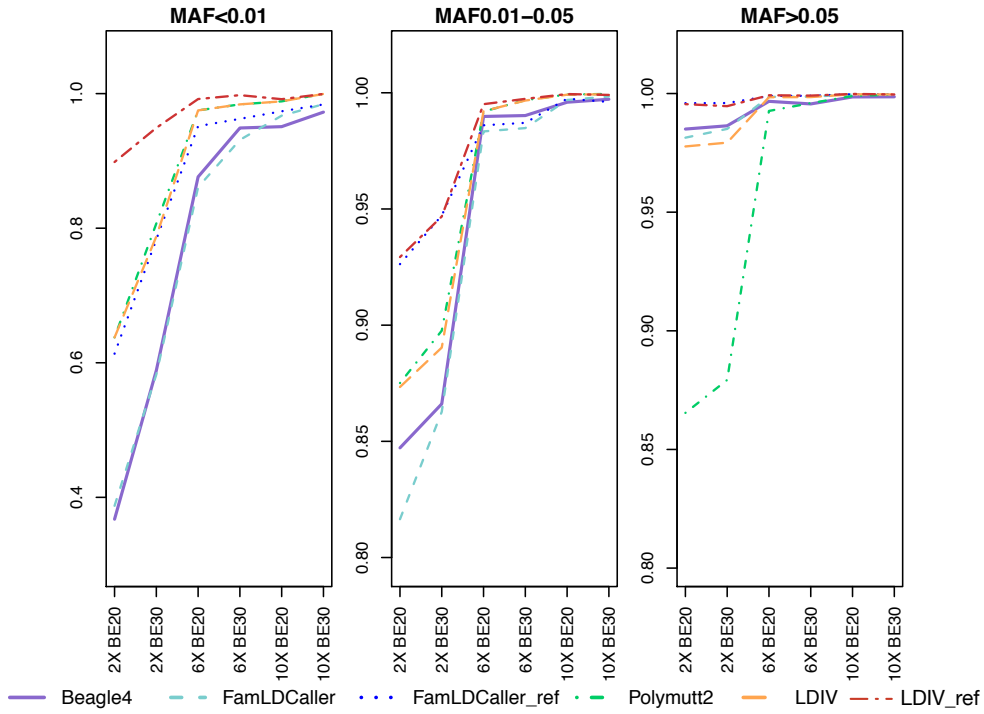


Figure 10: The genotyping accuracy on heterozygotes sites by MAF for nuclear family type 4.

Table 3: Mendelian error rate for family type 4

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	23.37	16.15	6.008	0	9.983	4.608
2X BE30	21.33	16.23	5.692	0	7.458	4.25
6X BE20	5.067	3.525	1.358	0	0.7167	0.6167
6X BE30	5.017	3.025	1.25	0	0.8333	0.5667
10X BE20	2.458	1	0.4583	0	0.1583	0.15
10X BE30	1.833	0.825	0.5417	0	0.275	0.3333

Table 4: Phasing accuracy for family type 4

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.99	0.972	0.9932	0.8785	0.954	0.9691
2X BE30	0.992	0.9773	0.9936	0.8966	0.9612	0.9694
6X BE20	0.9985	0.9964	0.9987	0.9562	0.9931	0.9692
6X BE30	0.9987	0.9976	0.9989	0.9771	0.9956	0.9698
10X BE20	0.9994	0.9993	0.9994	0.9832	0.9988	0.9774
10X BE30	0.9995	0.9996	0.9996	0.9941	0.9991	0.9806

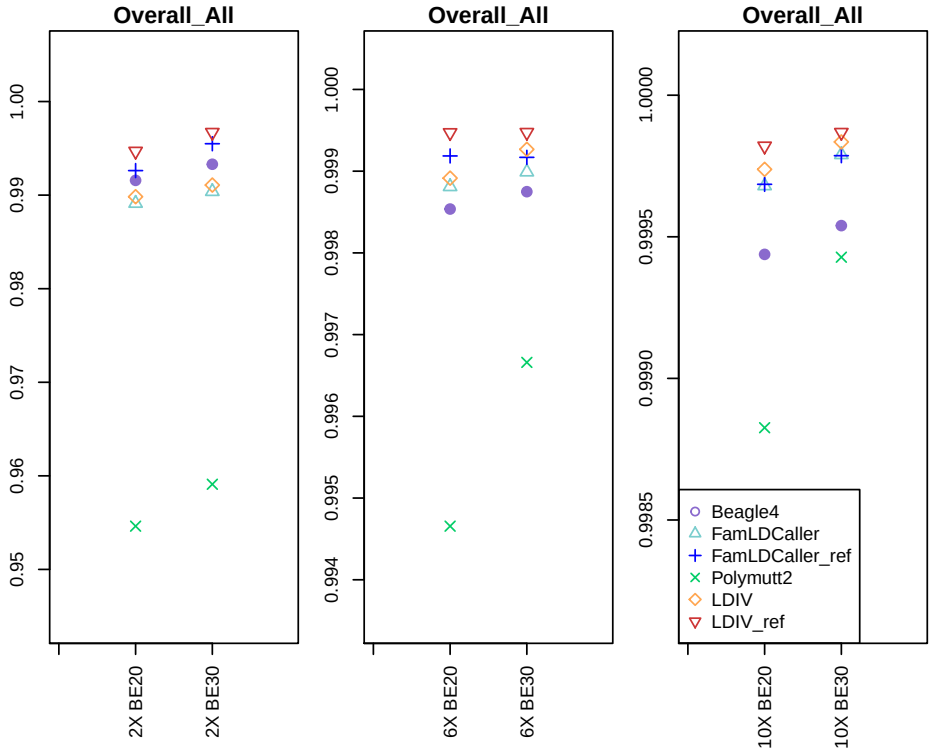


Figure 11: The overall genotyping accuracy for nuclear family type 5.

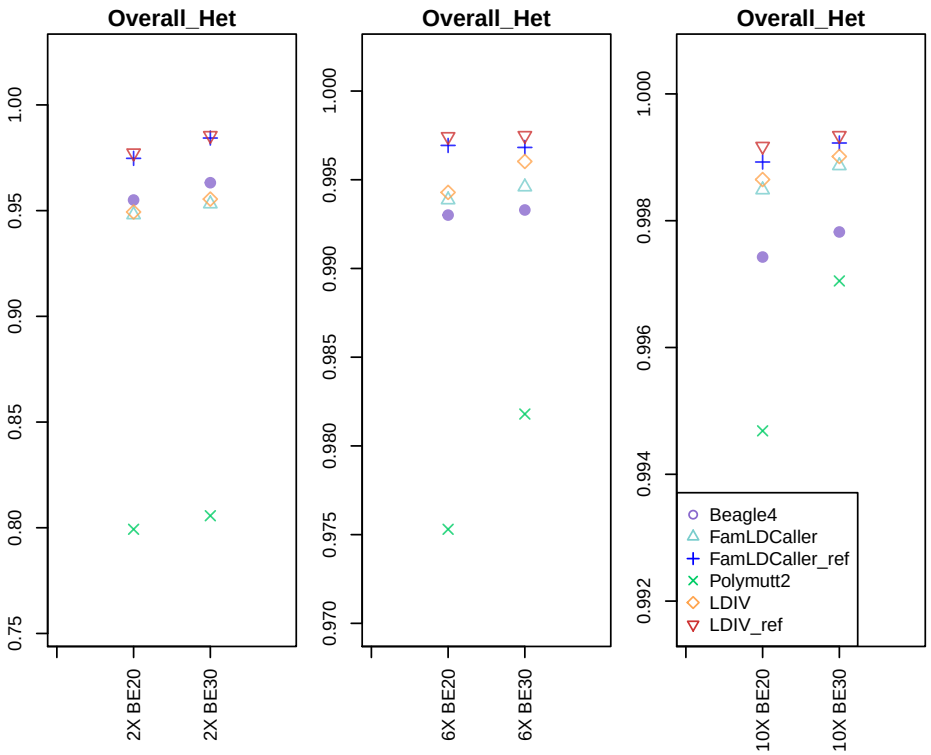


Figure 12: The genotyping accuracy on heterozygote sites for nuclear family type 5.

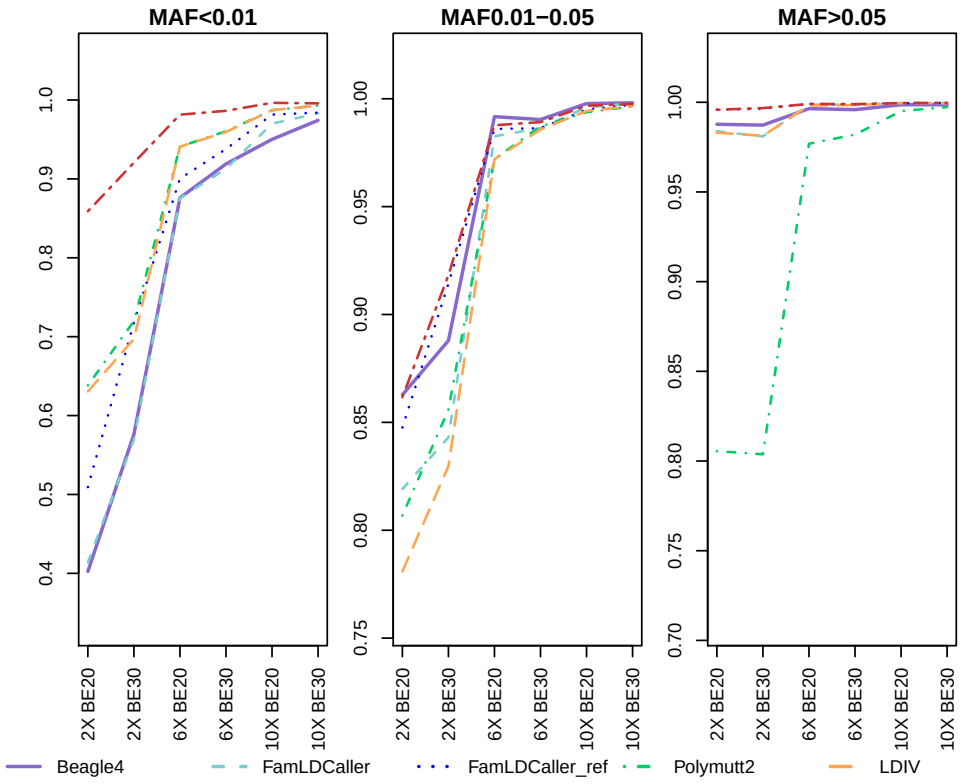


Figure 13: The genotyping accuracy on heterozygotes sites by MAF for nuclear family type 5.

Table 5: Mendelian error rate for family type 5

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	21.84	13.11	7.372	0	10.98	6.856
2X BE30	19.52	13.41	4.789	0	9.6	4.511
6X BE20	5.75	1.972	1.161	0	0.7444	0.7444
6X BE30	4.95	1.644	0.7611	0	0.7444	0.7056
10X BE20	2.233	0.55	0.3389	0	0.06667	0.1889
10X BE30	1.728	0.3556	0.1444	0	0.07222	0.1444

Table 6: Phasing accuracy for family type 5

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.9916	0.9705	0.9897	0.8612	0.9598	0.9703
2X BE30	0.9933	0.9726	0.9925	0.8563	0.9651	0.9748
6X BE20	0.9985	0.9961	0.9983	0.9767	0.9917	0.9812
6X BE30	0.9987	0.9952	0.998	0.9834	0.9933	0.9731
10X BE20	0.9994	0.9991	0.9994	0.988	0.9974	0.9811
10X BE30	0.9995	0.9988	0.9996	0.997	0.9983	0.9781

are established purely based on the inferred inheritance matrices like Polymutt2 will have no Mendelian error rate by theory (Table 3, Table 5, Supplementary Table 13). And since LDIV is built partially on inferred inheritance matrices, it has an edge over methods that ignored inheritance structures like Beagle4. And since FamLDCaller only consider trio family substructure inside a more complicated family, its performance will be better than Beagle4, yet not as good as LDIV. And lastly, we also see that including outside phased reference files will lower the Mendelian error rates. An example to illustrate all above observations is that in family type 4, with sequencing depth 2X and per-base error rate 0.01, Beagle 4 has most Mendelian error, 23.37 per individual, while Polymutt2 has 0. LDIV performed the best among LD aware methods, with 9.983 Mendelian errors per individual without reference, and 4.608 with reference; FamLDCaller performed third best, with 16.15, and 6.008 errors per sample without, and with reference, respectively.

3.4 REAL STUDY

3.4.1 Data description and preprocessing

We used the whole exome sequencing (WES) data collected in Kartini Clinic in Portland Oregon. The sequencing was conducted by capture kit Agilent version 4.0 on HiSeq 2000, funded by Klarman Family Fundation. The samples are from anorexia cohort with all anorexia patients and family members sequenced, consisting 74 individuals within 26 pedigrees. The family structures vary from unrelated individual, trio nuclear family, to nuclear family with 4 offspring. The average sequencing depth is $\sim 15X$. We take chromosome 22 for the purpose of demonstration in our real study, consisting of 80k variants. We followed the GATK best-practice procedure for all data preprocessing. Firstly, among the variants, only those that are deeply sequenced are taken as the ground truth for future comparison. To be specific, the original processed and genotyped data in variant calling format are filtered by minimum average sequencing depth greater than 20X, with 5k variants left, and their genotype calls as truth. Secondly, we do random subsampling on the original BAM files, with subsample

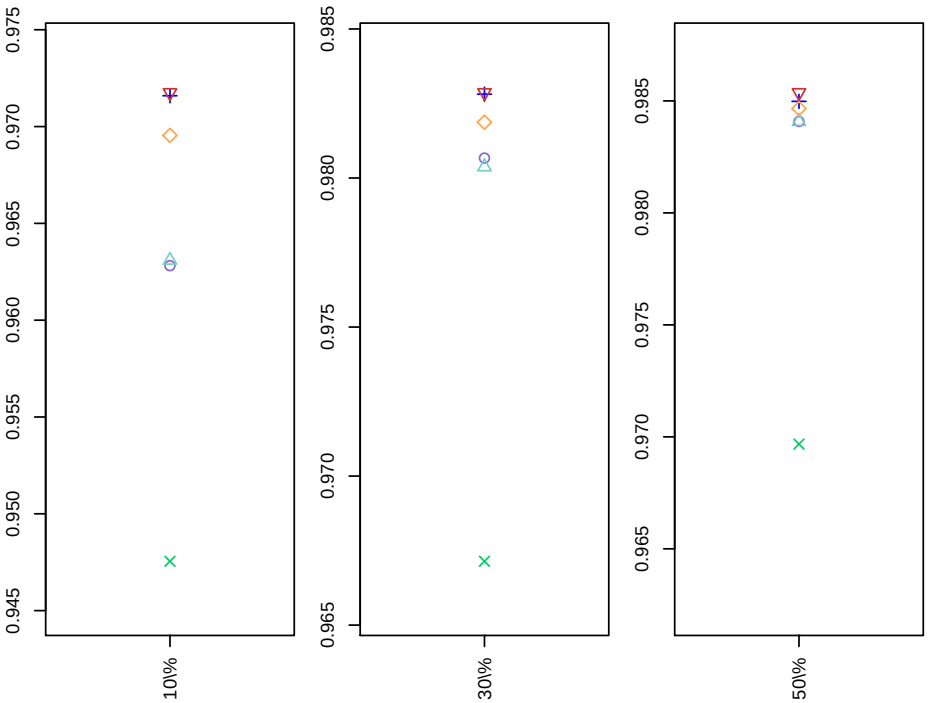


Figure 14: The overall genotyping accuracy for real data.

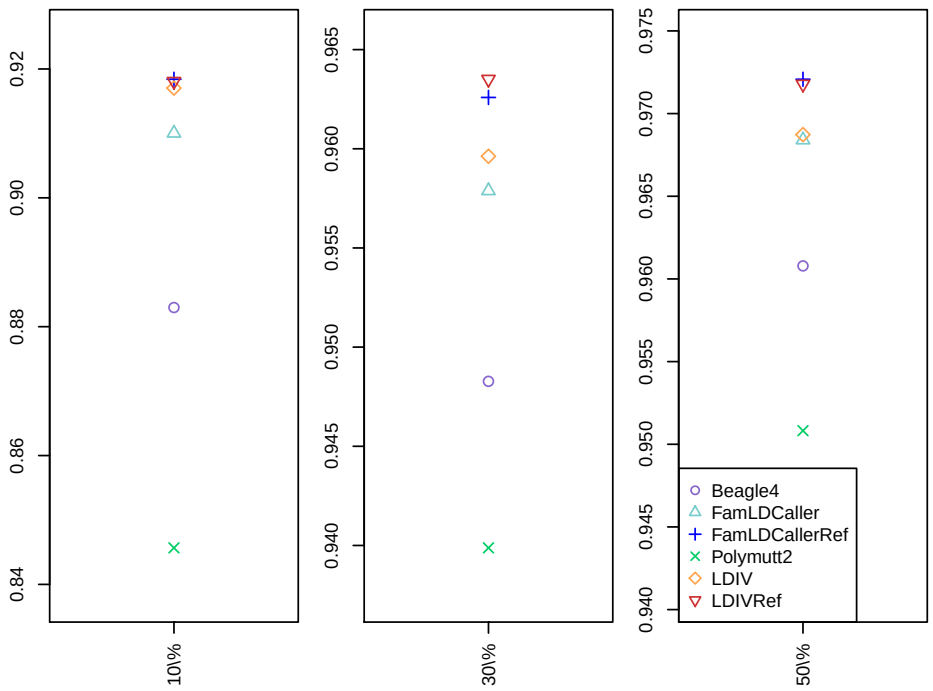


Figure 15: The genotyping accuracy on heterozygote sites for real data.

Table 7: Mendelian error rate (per family) for real data

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
10%	455.9	275.8	206.6	0	338.3	190.1
30%	1105	864.4	464.7	0	1039	430.5
50%	1929	1441	666.8	0	1755	633.9

proportion of 10%, 30% and 50%, mimicking shallower sequencing experiments, resulting in subsampled datasets with average sequencing depth of 4X, 6X, and 9X, respectively. The subsampled VCF files are then used as input for each methods we aim to compare as in simulation studies. Similarly, benchmarking criteria are overall genotyping accuracies, genotyping accuracies for heterozygotes only, and Mendelian error rates. We didn't phase original data, so no phasing accuracy is compared in real study. And by the sample size limitation, the number of rare variants (MAF $< 1\%$) after filtering by minimum average depth or subsampling are not sufficient for comparison, so we didn't separate variants by MAF in our evaluation. The reference panel in this study is taken from chromosome 22 of 1000 Genome Project Phase 3 data, with 2504 individuals in 26 populations.

3.4.2 Real study results

3.4.2.1 Genotyping accuracies We first compare the genotyping accuracy on the overlapped variants in truth and subsampled data for methods as in simulation studies. As expected, the genotyping accuracies increase as sequencing depth increases, which is induced by the increasing subsampling percentage in this study (Figure 14, 15). And also we see, when outside reference panel is used, generally the genotyping accuracy for both all variants and for heterozygotes only will increase. Furthermore, similar to simulation studies, we see that with or without reference files, LDIV almost always outperform other tools in terms of genotype calling accuracies. The main improve in genotyping occurs when comparing overall genotype calling accuracies, where for example for 4X sequencing depth data, LDIV decreased genotype calling error rates by 23.3%, comparing the second best methods, FamLDCaller.

3.4.2.2 Mendelian error rates We then proceed to compare Mendelian error rates between methods. We found that LDIV is generally retaining decent Mendelian error rate. For example for 4X data, LDIV without reference panel has 455.9 Mendelian errors per family, which is more than FamLDCaller (275.8), but less than Beagle4 (338.3); while with reference panels, LDIV has the least Mendelian errors (190.1) among LD-based methods

(all but Polymutt2), while FamLDCaller has 206.6. Generally we found, among LD-based methods, without reference panels, LDIV has less Mendelian errors than Beagle4, but more than FamLDCaller. And as expected, IBD-based method, Polymutt2, achieves absolute 0.

3.5 DISCUSSION AND CONCLUSION

We proposed a statistical and computational method, namely LDIV, to improve genotype calling accuracy in family-based next-generation sequencing data. Sequencing family data is crucial in uncovering association patterns of rare inheritable diseases. Compared to current genotype calling methods for family data, LDIV is advantageous in that it considers both linkage disequilibrium patterns and Mendelian inheritance patterns. The LD patterns are estimated either from samples within studies, or references including more deeply-sequenced or confidently-called samples from larger cohorts, like 1000 Genome projects. On the other hand, IBD patterns are estimated within each family. The inferences of both sets of patterns are effectively done in hidden Markov models, making LDIV a two-hidden-Markov-models-powered method. One possible improvement in the future could be sampling inheritance vectors by their posterior probability and do majority vote to summarize IBD pattern, instead of using the one and most likely inheritance vector I^{best}. This is expected to make the outcome more robust against uncertainty from IBD pattern inference.

LDIV with joint likelihood takes into account of both LD and IBD information, and by balancing the two information, we are able to reach a higher genotype calling accuracy across different rarity, with different family structures, and with or without external reference genomes. In simulation studies, for some very informative families (Figure 8), in some encouraging scenarios, LDIV will reduce genotype calling error rates comparing to the second best method, up to 46.2% without reference, and up to 62.5% with reference. Although one may argue that the accuracies are already high for both methods, when calling SNPs across genome, this error rate will induce up to 10,000 wrongly called genotypes, considering the total 10 million SNPs in human genome. And LDIV performs the best when family structure is informative, i.e., with many siblings that could help increase confidence in the inference

of inheritance patterns and share sequencing reads information across individuals, as well as when reference haplotypes are available. Actually, the use of external reference is generally encouraged in LDIV to improve genotype calling for variants of moderate to high minor allele frequencies. Meanwhile, LDIV will weigh on more informative families to help genotype calls on rare loci.

Similar to simulation studies, in real analysis, we found that LDIV is generally able to achieve the highest genotyping accuracy for both all variants and for heterozygotes loci only. Although the improvement is not as astonishing as in simulation studies where informative family structures are used, the improvement in reducing genotyping discordance rate could be up to 23.3% comparing LDIV to the second best method. And this advantage is especially strong when sequencing depth is low, meaning that LDIV is able to meet the need of handling data from cost-effective sequencing experiments.

LDIV will also balance Mendelian inheritance and phasing accuracy. As for simulation studies, the Mendelian error rate of LDIV is decently low, although not as low as pure Mendelian-inheritance-based methods like Polymutt2, which sacrifice genotype calling accuracy to achieve absolute Mendelian accuracy. As for phasing accuracy, LDIV is keeping a fairly low switching error rates throughout all simulation scenarios in simulation studies. We should also note that comparing phasing accuracy is disadvantageous for methods having higher genotype calling accuracies, in this case, LDIV. Because the loci that are incorrectly called will not incur a switch error, but only the loci correctly called will. Furthermore, in real studies, LDIV will outperform Beagle4 without reference panel, and outperform all other LD-based methods when reference panel is provided. Interestingly we found that the high number of Mendelian error rates could be partially, apart from the small percentage of multiple-sibling families in this dataset, responsible for the low genotype calling accuracy of Polymutt2, which strictly follows Mendelian's Law. And therefore, by balancing family information as well as LD patterns, LDIV is able to achieve both high genotyping accuracy and low Mendelian error rates.

In addition to the improvement in genotype calls by leveraging family and LD information, several other methods are proposed to approach the optimization of genotype calling in different angles. PhredEM is a method using EM algorithm to improve estimation accuracy

of base call error rates13. Other methods seek to reduce bias in allele frequency estimation, by either maximum likelihood method, or empirical Bayesian method. Similar to GATK19, those methods can be used in bundle with LDIV to achieve overall best genotype calls. Either these methods could use processed sequencing reads to make first round genotype calls, which will be later refined by LDIV; or they may serve as a correction tools to provide less biased estimations of base calling error rates and allele frequencies to LDIV for optimized inferences.

LDIV is implemented efficiently in C++. And by using dynamic programming techniques in inferences of hidden Markov models, LDIV is decently fast and feasible in terms of memory usage and computation time on a regular server, even for whole exome and even whole genome sequencing data. To better meet the need of smaller lab without rich computation power, especially with limited memory, we also provided a fully automated pipeline for splitting variants to do genotype calls using LDIV according to available memory. LDIV is freely available on both Github and https://sites.google.com/site/arkzhf9/home/ldiv, together with testing datasets, relevant pipelines, and user manuals with example starting guide that is easy to follow for users with basic Linux command line knowledge.

4.0 COMPARATIVE PATHWAY INTEGRATOR: A FRAMEWORK OF META-ANALYTIC INTEGRATION OF MULTIPLE TRANSCRIPTOMIC STUDIES FOR CONSENSUAL AND DIFFERENTIAL PATHWAY ANALYSIS

4.1 INTRODUCTION

In a typical transcriptomic study, a set of candidate genes associated with diseases or other outcomes are first identified through differential expression analysis. Then, to gain more insight into the underlying biological mechanism, pathway analysis (a.k.a. gene set analysis) is usually applied to pursue the functional annotation of the candidate biomarker list. The rationale behind pathway analysis is to determine whether the detected biomarkers are enriched in pre-defined biological functional domains. These functional domains might come from one of the publicly available databases such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). Three main categories of pathway analysis methods have been developed in the past decade. The first method called "over-representation analysis" considers biomarkers at a certain DE evidence cutoff and statistically evaluate the fraction of DE genes in a particular pathway found among the background genes. Without a hard threshold, the second category "functional class scoring" takes the DE evidence scores of all genes in a pathway into account and aggregates them into a single pathway-specific statistics. The third category "pathway topology" further incorporates the information of inter-gene interaction and their cellular location in addition to the pathway database (Khatri and Butte, 2012).

Many transcriptomic datasets have been generated with the rapid advances of high-throughput genomic technologies in the past decade. Meta-analysis, a set of statistical

methods for combining multiple studies of a related hypothesis, has thus become popular. Few methods have been developed for the pathway meta-analysis so far (Huang et al., 2009). Shen and Tseng (2010) developed two approaches of meta-analysis for pathway enrichment by combining DE evidence at the gene level (MAPE_G) or at the pathway level (MAPE_P). MAPE_P, which performed meta-analysis on the pathway level to generate a combined evidence score for each pathway. When multiple datasets for a common hypothesis are available under different conditions (e.g., tissues or labs), it might also be interesting to detect pathways enriched consistently under all conditions (consensual pathways) and pathways enriched solely in one condition but not in the others (differential pathways). One naïve way is to identify the enriched pathways in each study individually and manually check whether a certain pathway is enriched in one or multiple studies. To the best of our knowledge, there is currently no available statistical tool that can achieve this goal systematically.

One issue emerges with pathway analysis is the pathway redundancy. The large amount of individual pathways identified can hardly infer the underlying biology directly due to this issue. This kind of redundancy typically occurs in a regular pathway analysis since different pathways may include many overlapping genes. Toolkit DAVID resolved this issue by clustering pathways based on a kappa statistics representing the pathway similarity (Cohen, 1968). But the users still need to manually inspect every pathway in a cluster. Furthermore, due to the long and vague pathway description, users can barely make solid conclusions from the results.

In light of this, we proposed a framework of meta-analytical integration of multiple transcriptomic studies for consensual and differential pathway analysis, wrapped in a tool named Comparative Pathway Integrator (CPI). Our tool incorporates 24 databases, such as MSigDB, GO, KEGG, or user-defined gene set lists, as reference of pathway analysis. In order to identify both commonly and study-specifically enriched pathways, we applied the adaptively weighted Fisher's method, which is originally developed to combine p-values from multiple genomic studies for detecting homogeneous and heterogeneous differentially expressed genes. Clustering analysis based on the overlapping genes among enriched pathways is applied to remove the level of pathway redundancy. Subsequently, we developed a text mining algorithm to automate the annotation of pathway clusters by extracting key-

words from pathway descriptions, which also offers more statistically valid summarization compared to leaving user exploring each pathway in a cluster manually and heuristically. Lastly, CPI visualize the findings and provide users both text and graphical outputs for intuitive while statistical solid presentation and easy interpretation. An R GUI package CPI is available online to perform the analysis.

4.2 MATERIALS AND METHODS

4.2.1 Workflow of Comparative Pathway Integrator (CPI)

CPI is a comprehensive tool incorporating several widely accepted mature methods as well as some novel algorithms/approaches. It is mainly composed of three steps (see Figure 16). The first step is meta-analytic pathway analysis, which include pathway enrichment analysis and meta analysis. This step partially resembles the work of R package MetaPath. While MetaPath focuses on detecting consensus expressed pathways, CPI will detect both consensual and differentially expressed pathways, providing extra information on how the patterns of pathway enrichment differ across studies. The second step is pathway clustering. This step aims to reduce the redundancy of the pathways from normally hundreds of enriched pathways to a few clusters of pathways. The results are more succinct and interpretable. The third step is text mining based on the pathway names and descriptions to find keywords characterizing the overall information of the cluster. A permutation-based statistical test is provided to assess if a specific keyword significantly appear more than by chance. Without this step, it would be difficult to identify the representative characteristics for each cluster and users may still need to eyeball all the pathways in results since the step of clustering of the pathways does not reduce the total number of pathways. Finally we have both graphical output, containing heatmaps and topGO (optional) (Alexa and Rahnenfuhrer, 2010), and spreadsheet output, containing p-value matrices of pathways and pathway clustering details including gene composition and keywords.

4.2.2 Meta-analytic pathway analysis

Compared with differentially-expressed gene analysis, pathway analysis gives more biological insight and is more systematic and comprehensive. In CPI, we adopted method of over representation analysis (ORA) when users input gene list and corresponding p-values from each study. For users who preferred other types of pathway analysis, we also accepts lists of significant pathways and the corresponding p-value as user input.. Given pathway enrichment results, we perform Adaptively-weighted Fisher (AW Fisher) method (Li et al., 2011) as meta-analysis, to identify pathways significant in one or more studies/conditions. AW Fisher method not only increases the statistical power , but also gives a binary weight for each pathway, indicating the significance of the pathway in each study/condition. Given a user-specified q-value cutoff, we have a list of significant pathways, with some of them are commonly significant across studies/conditions while some of them are significant in specific study/condition.

4.2.3 Pathway clustering for reducing redundancy and enhancing interpretation

Because of the nature of pathways (e.g., hierarchy structure), many genes are shared among different pathways. This redundancy often reduces the interpretability of the result from pathway enrichment analysis. In CPI, we perform pathway clustering method to reduce the redundancy among pathways. The similarity between different pathways is calculated based on Kappa statistics, which depends on how many genes are mutual and exclusive among those pathways. The kappa statistics represents the distance between two pathways based on the genes composing each pathway. A distance matrix is then defined based 1-similarity matrix, composing the distance between each pair of pathways. We provide options for users to choose either consensus clustering or fuzzy cluster to perform the clustering analysis. When consensus clustering is chosen in this step, follow the original consensus clustering method, an elbow plot and consensus CDF plot are generated to help users decide the total number of clusters

For most clustering algorithm including the consensus clustering we adopted, each pathway will be forced to cluster into one group of pathways, no matter it's scattered or not

compared to the pathways within the same cluster. We allow scattered pathways to form singletons, when its gene composition is different from other pathways, to avoid adding noise to other existing clusters. To improve the tightness of the cluster, we further calculated the silhouette width, a measure of how tightly each pathway is grouped in its cluster, and removed the scattered pathways with low silhouette width. The removing cutoff for silhouette width is estimated empirically based on its distribution from our multi-disease dataset. For those singletons identified, we collect them into another cluster instead of filtering them out because those pathways might be interesting in terms of unique genes composition.

4.2.4 Text mining for automated annotation of pathway clusters

4.2.4.1 Question description Reducing the redundancy of pathways by clustering per se gives limited summary of the pathways. Since the user usually need to go through most of the pathways in a cluster to grasp an idea and interpretation of the contents in the cluster, the interpretation is not quantitative and largely rely on the biological knowledge of the user. Therefore we need a more rigorous statistically meaningful representation and summary of the cluster.

The above goal is expressed here as the text mining for key words for each clusters. So the question becomes: which word appears more frequently in a certain cluster than they usually would? We will therefore treat these words as the keywords for that cluster. The key word is counted based on the number of pathways containing the word, rather than the frequency of the word appearing in all pathways in a cluster.

4.2.4.2 Pathway-word matrix We first extracted the unique words in description and term name of a pathway. Then we use R package Apache openNLP for tokenization, i.e. identify nouns, adjectives, prepositions and etc. Plural nouns are combined with nouns, words other than nouns and adjectives are filtered out because they are not informative enough to be keywords of the cluster. The common English words and the words that are found only in one pathway are filtered out for the same reason. Then we use R package SnowballC to find word stems, so that the adjectives are combined with their corresponding

nouns. The remaining is a binary matrix where each row being a word and each column being a pathway, with element $x_{ij} = 1$ indicating that the pathway j contains the word i.

4.2.4.3 Test statistics A simple strategy to test for the significance of a word in a cluster is by simple counting and conduct Fisher exact test. Yet we found this method to be less powerful and biological justifiable from real data analysis, because the words in the term name or a shorter description of a pathway are designed to be more representative than those in a full or longer description. We therefore decide to penalize on words found in longer description. We down-weighted the word count by assigning a score between 0 and 1 to each pathway j to indicate whether it contains word i:

$$x_{ij} = \begin{cases} 1 & \text{word } i \text{ appeared in term name of pathway } j, \\ exp(-\alpha|w_j|) & \text{word } i \text{ appeared only in description on pathway } j, \\ 0 & \text{otherwise} \end{cases} \tag{4.2.1}$$

Where $|w_j|$ is the number of unique words in description of pathway j; α is a parameter controlling the intensity of penalty. The greater the α is, the greater the penalty is on longer description. When α equals to 0, there is no penalty and our test simplifies to is equivalent to Fisher exact test.

And then we defined cluster score $T_i(C)$ to be the sum of scores of pathways in the cluster, i.e. for word i in a pathway cluster C, we have our test statistics:

$$T_i(C) = \sum_{j \in C} x_{ij}$$

4.2.4.4 Permutation test To test for the null hypothesis that a word is not enriched in a certain cluster, we adopt a permutation test. The basic argument of constructing the permutation distribution for the test statistics, under the null hypothesis, is that all words occur equally frequent across all clusters, including the unpermuted data. So for each word i in the tth permutation, pathways are randomly sampled to form subset S_t with the same cluster size as C. Test statistics $T_i(S_t)$ is recomputed at the end of each permutation. The

operation is then repeated for a large number of times (say, T times). At last, all $T_i(S_t)$ are ranked together with the original data $T_i(C)$. And the p-value could be calculated thereby, indicating how extremely frequent word i is seen in cluster C.

4.2.4.5 Graphical and Spreadsheet Output Suppose we already have the clustered pathways and the AW Fisher p-value for each pathway, in this final step we aim to help users understand the overall pattern of pathways significance better, by visualization approaches including heatmaps of p-value for each pathway under each condition, heatmap of kappa statistics for each pathway, and topGO graphs as an option if the input pathway database contains GO.

4.2.4.6 Implementation CPI is implemented in R with packages, tools and novel methods mentioned above. A standalone R version of CPI is ideal for users with basic knowledge of R and command line interface, and could take advantage of parallel computation on server. A GUI version CPI is programmed with the help of R GUI project. This version of CPI can be used by user of any level of programming proficiency while achieving the same results as standalone R version.

4.2.4.7 Datasets and databases We provide 21 homo sapiens pathway database including 14 pathway database from MsigDB, 2 database from Connectivity map, transcription factor target database JASPAR, Protein-Protein interaction database and 3 microRNA databases as options for enrichment analysis.

In addition, Gene Ontology, KEGG database for organism Mus musculus and Saccharomyces cerevisiae, JASPAR database for organism Mus musculus are provided.

4.3 RESULTS

4.3.0.1 Constructing the pathway-word matrix We first construct the pathway-word matrix using the filtering steps described in methods section. Among the total 24302

input pathways, we found 14251 unique words in term names and 34661 in full descriptions. Pooling the words yields 36957 unique words in total. Starting from 36957 unique words, we applied all filtering steps and trimmed the matrix down to 12949 unique words in 24302 pathways (Bird, 2006).

4.3.0.2 Justifying penalization in text mining

Setting the text mining permutation at 1000 times took a reasonable time for the whole analysis procedure: nine minutes. Our results also demonstrate the power of penalized permutation test over Fisher's exact test. Since Fisher's exact test treats words in description (can be up to 1500 words) the same as words in pathway names (usually less than 15 words), signals of common biological words in pathway names will be masked due to its high occurrence in pathway descriptions and some non-informative words in descriptions will be detected as keywords falsely by chance. For example, in our text mining results of the two tests, "ATP" in the mitochondrial ATP activity cluster was ranked low in the Fisher's exact test (rank = 19) while prioritized in permutation test (r=1). Some meaningless words such as "buolo" and "engelhardt" in a ribosome cluster was ranked high by Fisher's exact test (r = 4), but ranked low by penalized permutation test (r = 34). To conclude, for detecting ordinary keyword, penalized permutation test performs roughly the same as Fisher's exact test, however, for detecting keyword of frequently occurred biological vocabulary or filtering out strange jargons penalized permutation test is indeed powerful than Fisher's exact test.

4.3.0.3 Results for real data

To demonstrate its utility, we applied CPI using the default databases to analyze the transcriptome datasets of three psychiatric diseases of two prefrontal cortex layers. These datasets, provided by Dr. David A. Lewis' group, was used previously to compare post-mortem tissue dorsolateral prefrontal cortex (DLPFC) layer 3 and layer 5 pyramidal cells' gene expression level of bipolar disorder, major depressive and schizophrenia patients with matched healthy control (Reinhart, 2015).

By inputting the top 400 differentially expressed genes in each of the six datasets and default pathways containing 10 to 200 genes, top 100 major pathways of total 2187 pathways

were identified and clustered to 7 clusters with 31 pathways as singleton terms. Of seven clusters, three important clusters with insightful biological meaning are discussed below. The elbow plot and consensus CDF plot (Figure 17ab) indicate that a cluster number of 7 is reasonable. Because when the number of clusters is greater than 7, the relative change in area under CDF curve of elbow plot goes slows down (Figure 17a), and the consensus CDF flattens out (Figure 17b).

Based on the heatmap output shown in Figure 19, we found that pathways in cluster 1 are significantly altered in schizophrenia DLPFC layer 5. Cluster 2, 5 and 7 shows similar pattern, and are significantly enriched consensually in both schizophrenia DLPFC layer 3 and 5. Cluster 3 is enriched in the DLPFC layer 5 of two diseases: significantly in major depression disorder, and marginally, bipolar disease. Cluster 4 is solely enriched in the layer 3 of schizophrenia. Cluster 6 is enriched significantly in layer 3 of schizophrenia, and slightly in the layer 3 of schizophrenia and bipolar disease. And we also observed that the singletons on the bottom of the heatmap display no consensual nor differential pattern, as they are scattered pathways.

To further demonstrate the consensual and differential patterns of our results, we inspect the contents of cluster 2, 3 and 6 with the help of text mining. The information provided from these three pathways are diverse, even within a cluster. But those information could be condensed statistically into keywords, with our penalized text mining method.

Cluster 2 consists of 8 pathways: KEGG Huntington's disease, KEGG Alzheimer's disease, Reactome Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins, Reactome Respiratory electron transport, Reactome The citric acid (TCA) cycle and respiratory electron transport, KEGG Parkinson's disease, KEGG Oxidative phosphorylation, KEGG Cardiac muscle contraction. Some of the top keywords for cluster 2 under FDR cutoff 0.02 are: dysfunction, mitochondrial and ATP, with word count 3, 5 and 3, respectively.

Cluster 3 consists of 10 pathways: BioCarta Role of Erk5 in Neuronal Survival, BioCarta Human, Cytomegalovirus and Map Kinase Pathways, KEGG Fc epsilon RI signaling pathway, KEGG Bladder cancer, KEGG Thyroid cancer, GO:MF MAP kinase activity, Reactome DSCAM interactions, Reactome Signalling to ERKs, Reactome Signalling to RAS,

Reactome p38 MAPK events. The keywords under FDR cutoff 0.02 are: ERK2, MAPK and MEK, with word count 3, 3 and 2, respectively.

Cluster 6 consists of 16 pathways. Manually inspected the pathways, we found this cluster contains many ubiquitin-proteasome related genes. The text mining results consent with that observation by giving top representative keywords under FDR cutoff 0.02 as: proteasome, degradation and CDC6, with word count 3, 8 and 3, respectively.

4.4 DISCUSSION AND CONCLUSION

In this article, we explored the approaches for comparative meta-analytic pathway analysis, and developed an integrative platform for this purpose called "CPI". CPI reduces pathway redundancy to condense knowledge discovered from the results and also conducts text mining to provide statistically solid suggestions on interpreting results. Our method has three advantages as compared to previous methods addressing pathway meta-analysis. First, CPI explores consensual and differential expression pattern spontaneously in integrated pathway analysis. Second, CPI clusters pathways by the gene composition to reduce pathway redundancy. Third, CPI uses a statistically valid text mining method to interpret pathway analysis results. In addition, the penalized text mining algorithm by permutation test has shown the advantage over the standard test like Fisher's exact test based on real data analysis. We applied the tool to multiple psychiatric disorders transcriptomic data. The result identifies multiple pathway enrichment patterns relevant to previously confirmed as well as novel biological functions, such as mitochondrial ATP dysfunction in schizophrenia DLPFC layer 3, ubiquitin-proteasome system dysregulation in schizophrenia and bipolar disorder DLPFC layer 5 and altered MAPK/ERK signaling chain in major depression disorder DLPFC layer 5 (Arion et al., 2015).

CPI starts with input files that are either p-value matrices of gene from different studies, or the Fisher p-value of significant pathways from each studies. Then CPI conduct meta analysis using AW Fisher's method to identify commonly and study-specific pathway enrichment patterns. Consensus clustering is performed on the those significant pathways to reduce pathway redundancy. And in the end, text mining is performed for the pathway clusters to

automate the interpretation, extracted the keywords that could be otherwise overlooked in either manual inspection of the results, or a more naive text mining method, for instance, Fisher exact test. An example in our data is that, in pathway cluster 2 relating to mitochondrial ATP dysfunction, the keyword ATP is ranked by Fisher's exact test as 19 and thus ignored (when we only inspect top 10 significant keywords), but is prioritized to top 1 by our penalized text mining method. In the end of CPI, the output heatmap and spreadsheet containing significant pathway clusters, enrichment pattern, as well as the top significant keywords representing the content of each pathway cluster are presented as results in each CPI run.

We applied CPI to transcriptome datasets of three psychiatric diseases of two prefrontal cortex layers from Dr. David A. Lewis' lab. Analyzing the top 100 major pathways,our results found that the MAPK/ERK signaling chain was altered in major depressive disorder DLPFC layer 5 but not in layer 3. Reduced MAPK/ERK signaling in hippocampal area is associated with depressive behavior. It is unknown whether major depressive disorder is associated with altered MAPK/ERK signaling in dorsolateral prefrontal cortex. Our study demonstrated altered MAPK/ERK signaling in major depressive disorder DLPFC layer 5 but not in layer 3 or the other two diseases.It remains an open question how this alteration relates to major depression disorder symptoms. We also found that mitochondrial ATP dysfunction in DLPFC layer 3 and to a less extent in layer 5 of schizophrenia subjects, which is consistent with a previous transcriptomic study of schizophrenia and schizoaffective disorder in Dr. Lewis' lab. And our results further indicated that the difference in the degree of mitochondrial dysfunction across DLPFC layer 3 and layer 5 was due to different degrees of mitochondrial ATP dysfunction rather than other aspects of mitochondrial dysfunction. Another previous finding from Dr. Lewis's lab which associated ubiquitin-proteasome related genes to schizophrenia layer 5 is also validated by CPI. Furthermore, consensual results across different diseases and layers are shown by the significant alteration of ubiquitin-proteasome system not only in schizophrenia DLPFC layer 5, but also in schizophrenia and bipolar disorder DLPFC layer 3. Similar result from a blood based microarray investigation evidenced preliminarily ubiquitin-proteasome dysregulation in both schizophrenia and bipolar disorder. However, the DE genes contributing to ubiquitin-proteasome dysregulation are dif-

ferent in bipolar disorder and schizophrenia, which may need further biological investigation and interpretation.

CPI has three major steps. In the first step, users may input either p-value matrix of either genes or pathways for each study to start with. If p-values of genes are inputted, pathway analysis is applied first. Then user may select a q-value cutoff for significant pathways (Benjamini and Hochberg, 1995). And those pathways are passed down to meta analysis where AW Fisher is applied to discover consensually and differentially enriched pathways. We suggest 0.2 as our default cutoff, and the users may also choose their own cutoff according to their budget on the follow-up experiments and analysis. We suggest âas as our default cut-off, users can choose their own value according to different scenario. In the second step, the pathways are clustered using consensus clustering, with consensus CDF plot and elbow plot assisting users to choose the number of clusters (Monti et al., 2003). By default, a group of six clusters were assessed. Silhouette information is used to achieve cluster tightness by moving scattered pathways to singleton. The cutoff of 0.1 is selected using silhouette information (Rousseeuw, 1987) distribution from our multi-disease dataset, and results in a reasonable number of singletons (30 out of 100 pathways). In the third step, text mining, users may choose number of permutation iterations based on user's computation capacity and data size. The computational complexities of the CPI R standalone and GUI version are similar. Using one thread on iX windows system, with the input of 6 gene lists with total of 84 subjects and 44k genes and the default pathway reference database, setting permutation test iteration to be 1000, one run of CPI takes 9 minutes. Moreover, both versions can be executed in parallel to further speed up the analysis.

CPI has several limitations. Firstly, for single study pathway analysis, our tool only allows user to apply Fisher exact test, but it can be readily extend to include other methods such as KS test and GSEA (Subramanian, 2005). Secondly, our text mining algorithm rely on the descriptions provided by pathway databases. So for those pathway databases that do not provide descriptions, e.g. some pathways in KEGG, text mining algorithm loses advantages. Thirdly, computation time is not ignorable, especially in text mining step. For the real data we applied, each 1000 permutation iterations increase 5 minutes computation time.

In summary, CPI is a meta-analytic tool for discovering commonly expressed and study specific pattern in transcriptomic studies, that will also reduce pathway redundancy and conduct text mining to increase interpretability of the results. CPI is implemented in R, as well as RGUI, an R based graphical user interface. The RGUI version can be easily handled by users without programming knowledge.

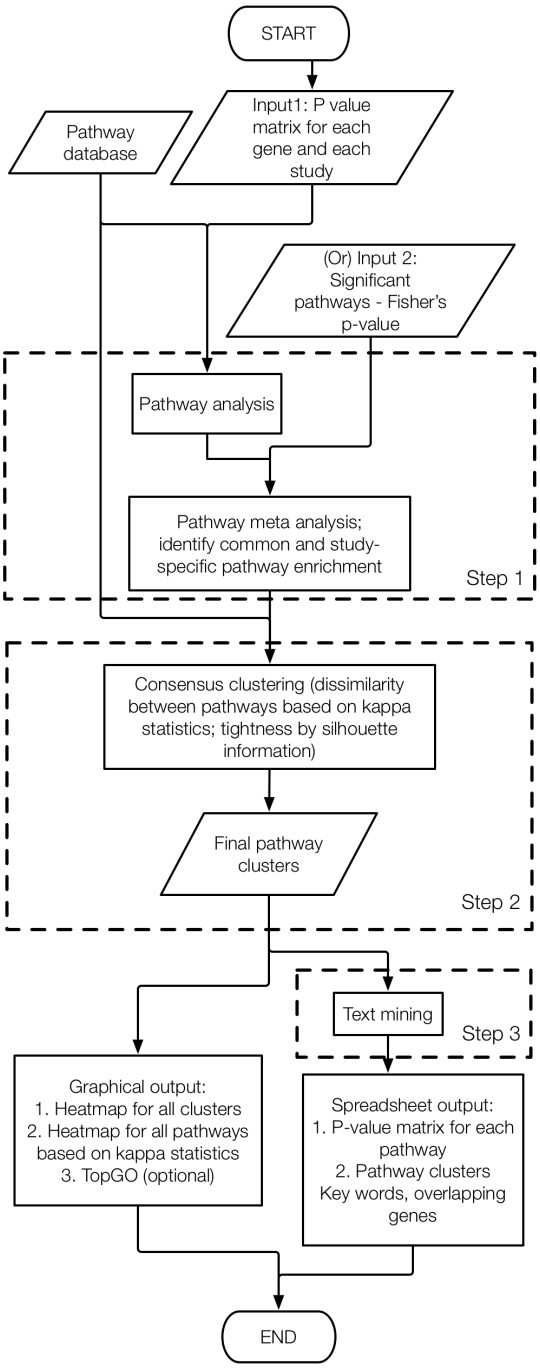


Figure 16: The workflow of CPI

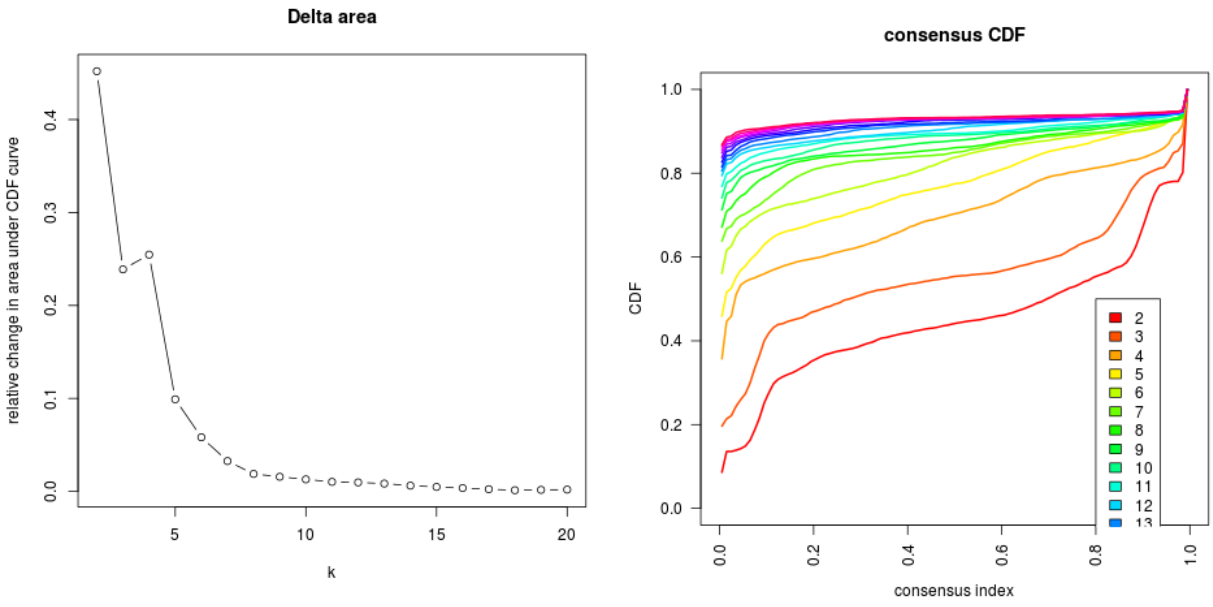


Figure 17: Plots used to assist decision on total cluster number.

(a) Elbow plot. **(b)** Consensus CDF plot.

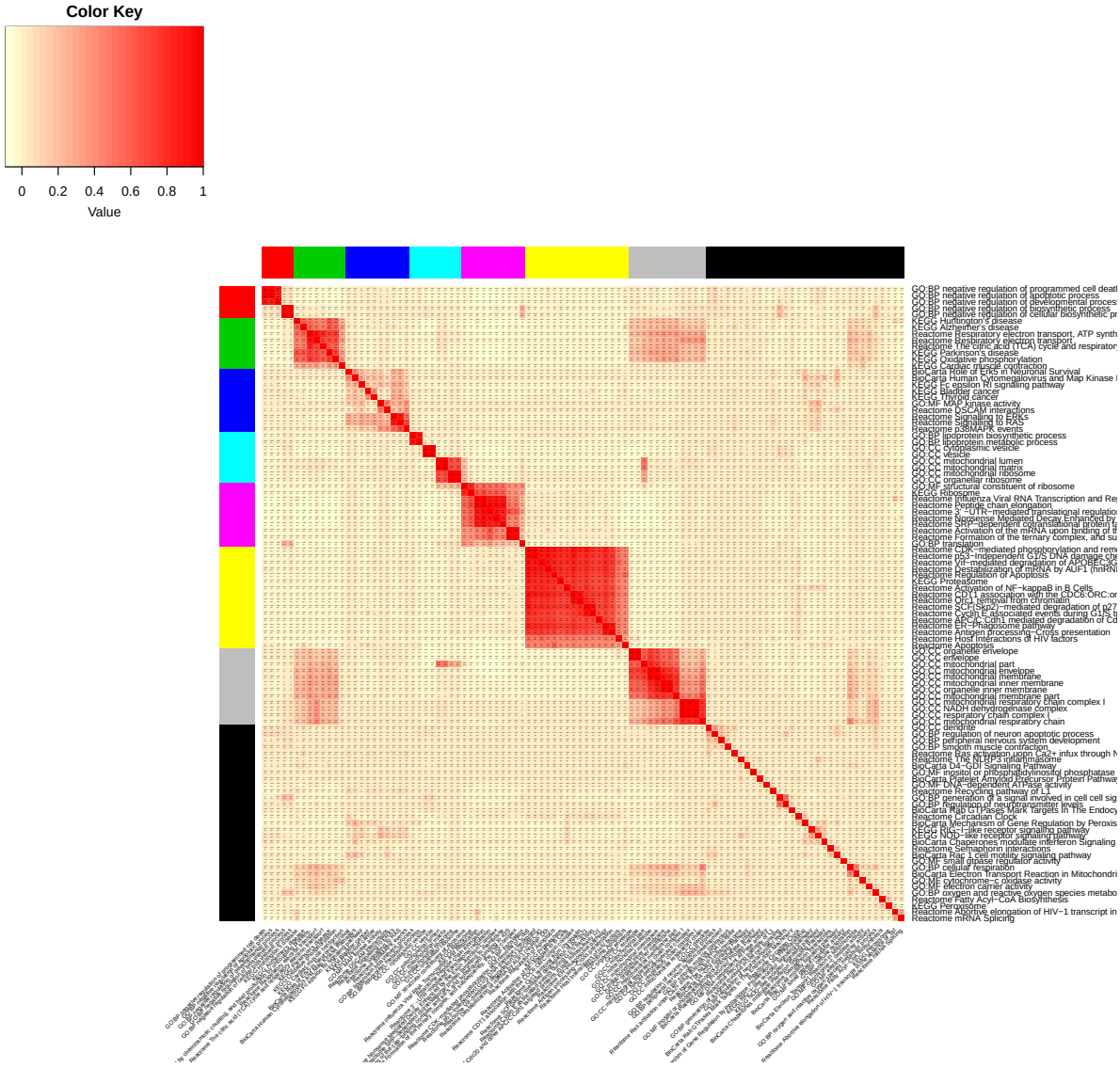


Figure 18: Heatmap of logged p-value of pathways for all clusters using four additional databases.

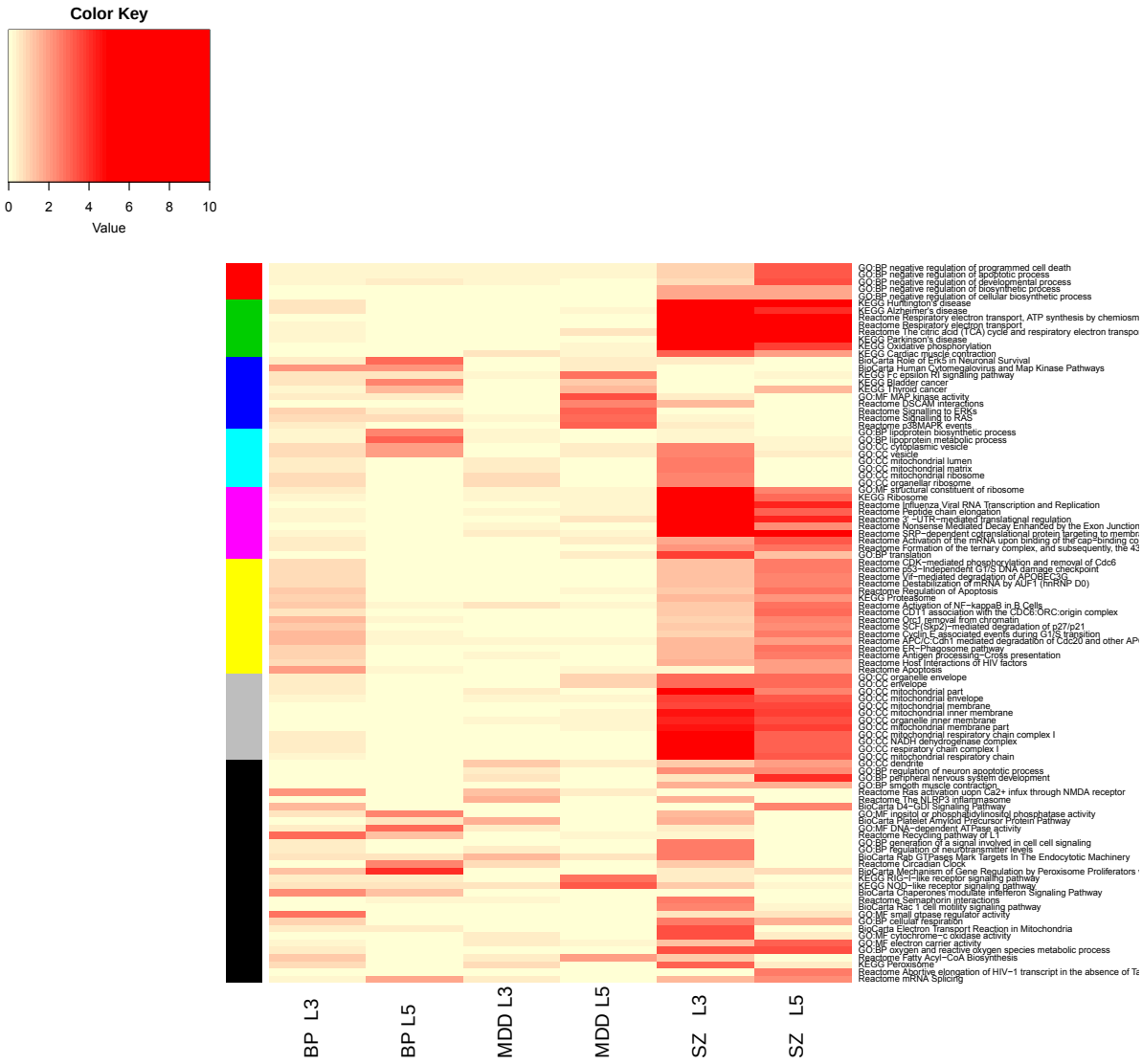


Figure 19: Heatmap of logged p-value of pathways for all clusters using default databases.

5.0 DISCUSSIONS AND FUTURE WORKS

5.1 DISCUSSION

The first paper proposed a Bayesian hierarchical model for multi-omics integration with missingness. With the modern computational technologies and algorithms, for example, MCMC Gibbs sampling and spike-and-slab priors, the construction and solution of complex Bayesian hierarchical model became feasible and popular. Especially when integrating heterogeneous data types like multiple omics data, Bayesian hierarchical model is promising in discovering the association between data types and selecting data-specific features. Our model takes advantage of this feature, and further incorporates missing data handling and self-paced learning decision making assistance.

The second paper proposed a novel computational algorithm to leverage both LD and IBD information to enhance genotype imputation. Both LD and IBD information have been proved to be a powerful in increasing genotyping accuracy. However, our method is the first to incorporate both information in one model, and provide fast and stable solution by its realization in hidden Markov model.

The third paper presented a meta-analytic pathway enrichment analysis framework. By applying AW-Fisher's method, pathway clustering, developing a text mining algorithm, and etc., this framework have several merits compared to currently available methods, including discovering differential and consensual pathways across studies simultaneously, reducing pathway redundancies and using text mining algorithm to facilitate scientists in discovering findings with more solid statistical support.

5.2 FUTURE WORKS

For the Bayesian multi-omics integration project, after the resubmission of the paper, I will consider using negative binomial distribution to better accommodate RNAseq gene expression count data when Gaussian assumption is violated. Futuremore, the model can be extended to new types of omics data, for instance, single cell sequencing (scRNASeq) data. And lastly, I will the statistical modeling under more complex missing patterns under missing not at random assumption.

Google Brain Team (DePristo and Poplin, 2017) recently proposed a deep neural network pipeline named DeepVariant for high-throughput sequencing data analysis. It's variant calling function could be a potential competitor to our LDIV method proposed in the second paper. In addition, although tested in real data simulation, the method in second paper also need to be used in real studies to support its performances.

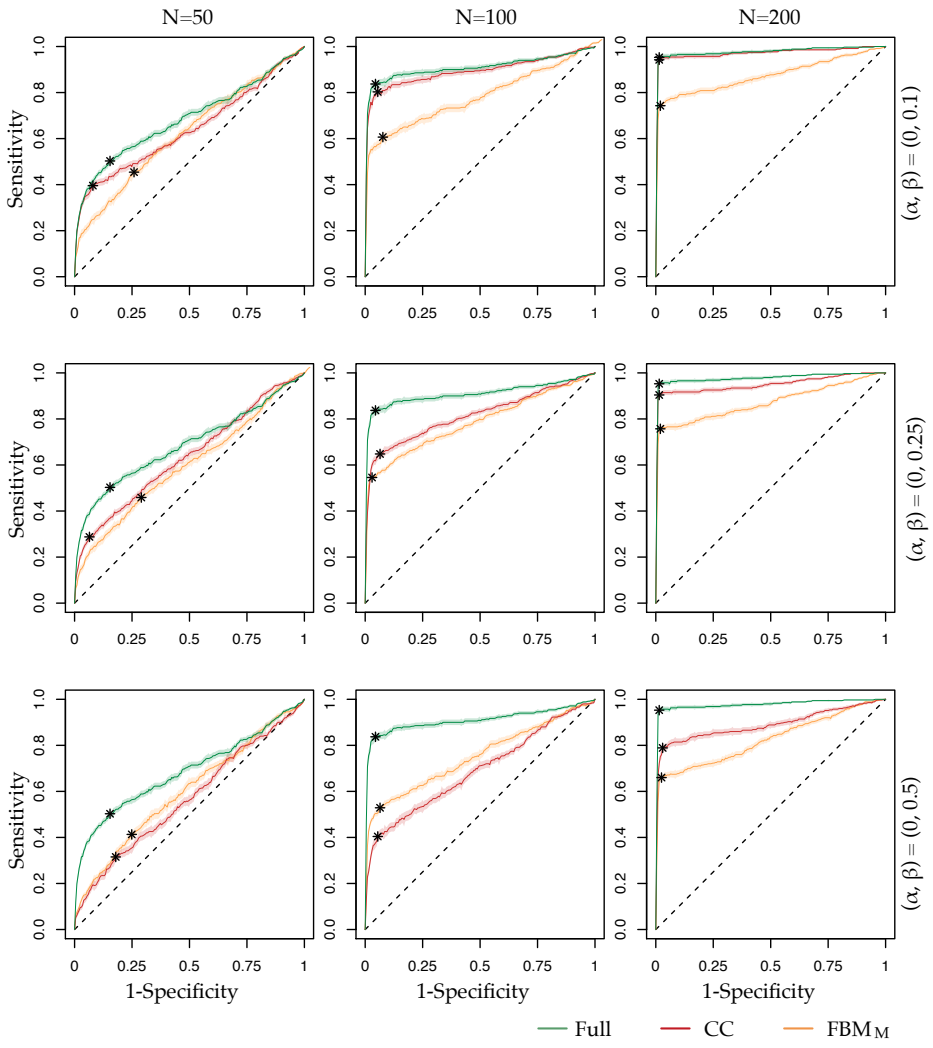
Furthermore, I am also open to explore and analyze new high-throughput data types in bioinformatics, and integrating them with currently available data types. In addition, I am interested in developing statistical and computational genomics methods and software to address pathology questions in cancer and complex disease studies. Lastly, I am interested to serve for bridging recent advances in computer science and machine learning with clinical and biological realms.

APPENDIX A

APPENDIX FOR MULTI-OMICS BAYESIAN MODEL WITH MISSINGNESS

A.1 SUPPLEMENTARY MATERIALS

a)



b)

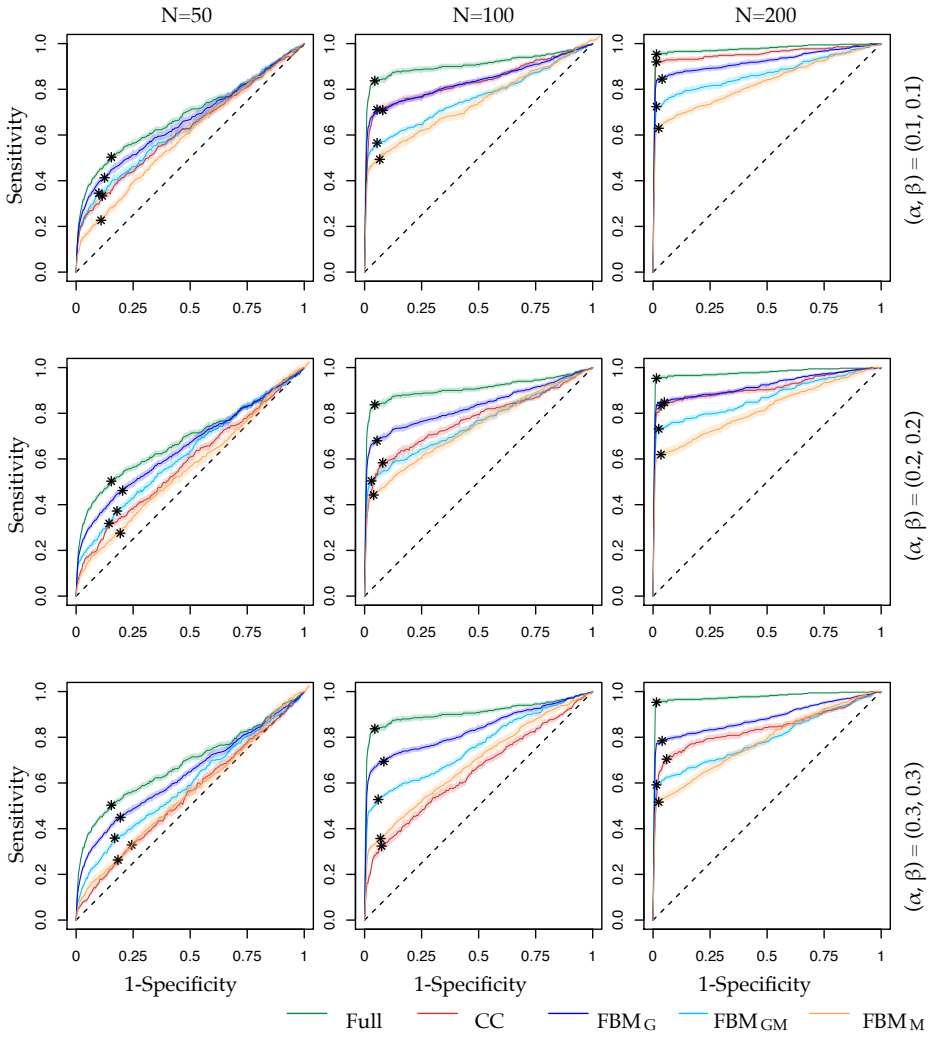
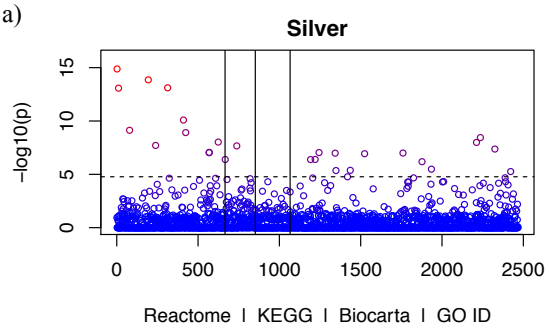


Figure 20: ROC curves for feature selection comparisons for Scenario II and III.

a-b) are the ROC curves for feature selection comparison for Scenario II and III following Figure 3. Scenario II: $\alpha = 0, \beta \neq 0$, Scenario III: $\alpha \neq 0, \beta \neq 0$. Star on each ROC curve is the point with maximum Youden index.

a)



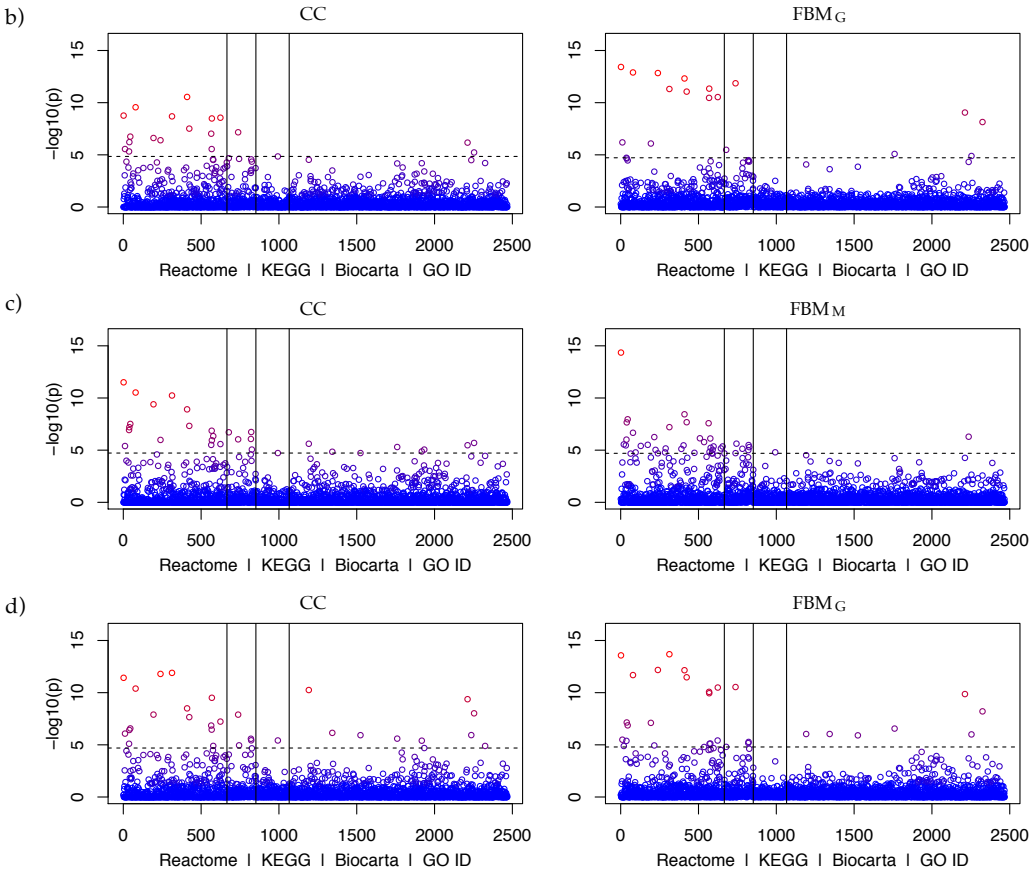


Figure 21: Manhattan plot for pathway analysis.

a) Manhattan plot for pathway analysis with full asthma dataset. b) Manhattan plot for pathway analysis with 50% samples missing gene expressions. c) Manhattan plot for pathway analysis with 50% samples missing methylation. d) Manhattan plot for pathway analysis with 20% samples missing gene expressions and 20% samples missing methylation, but no sample miss both gene expression and methylation at the same time.

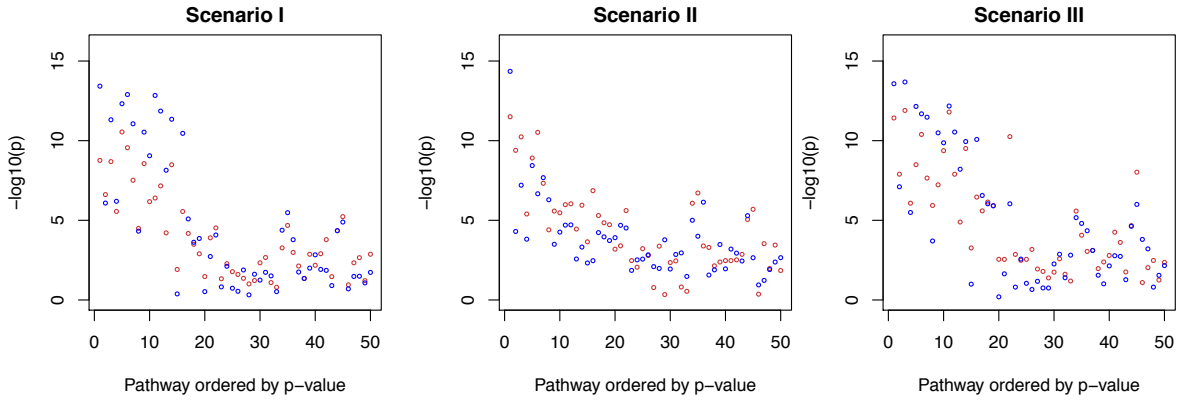


Figure 22: Pathway analysis similarity for pathway ordered by p-value.

X-axis is top 50 pathway index ordered by p-value from full data; y-axis is $-log_{10}(p)$ value of CC in red versus FBM_G (Scenario I and Scenario III) or FBM_M (Scenario II).

A.2 NON-INFORMATIVE PRIOR STRUCTURES AND MCMC GIBBS SAMPLER

A.2.1 Non-informative prior structures

We use $D = \{\mathbf{Y}, \mathbf{M_{obs}}, \mathbf{G_{obs}}, \mathbf{C}\}$ to denote the fully observed data, $\mathcal{M} = \{\gamma^M, \gamma^{\overline{M}}, \gamma^C, \Omega, \sigma, \sigma_k, \mathbf{G}_{mis}, \mathbf{M}_{mis}, \tau_k^M, \tau_k^{\overline{M}}, \tau_j^\omega, \mathbf{I}_k^M, \mathbf{I}_k^{\overline{M}}, I_j^\omega, \pi^M, \pi^{\overline{M}}, \pi^\omega\}$ to denote the full parameter sets, among which we are interested in $\gamma^M, \gamma^{\overline{M}}, \gamma^C, \Omega\}$. The full Bayesian Hierarchical model can be generally written as:

$$P(\mathcal{M}|D) \propto P(\mathcal{M})P(D|\mathcal{M})$$

We starting laying out our prior constructions with our full Bayesian model (with $\mathbf{G}^M$ equivalently substituted by $\mathbf{G} - \mathbf{M\Omega}$):

$$[Y|\mathbf{C}, \mathbf{G}, \mathbf{M}; \mathcal{M}] = \mathbf{C}\gamma^C + (\mathbf{M\Omega})\gamma^M + (\mathbf{G} - \mathbf{M\Omega})\gamma^{\overline{M}} + \epsilon$$

$$[\mathbf{g}_k|\cdot] = \mathbf{M}_{\mathcal{J}_k}\omega_k + \mathbf{g}_k^{\overline{M}}, \mathbf{g}_k^{\overline{M}} \sim MN_N(0, \sigma_k^2\mathbf{I}_{N\times N})$$

$$\epsilon \sim MN_N(0, \sigma^2\mathbf{I}_{N\times N})$$

where $\mathcal{J}_k$ is the set of all $\mathbf{m}_{j_0}$ within the promoter of kth gene, and $\mathcal{J}$ is the length of $\mathcal{J}_k$.

Priors on other parameters are relatively straightforward:

$$\left[\gamma^C\right] \sim MN_L(0, 10^6\mathbf{I}_{L\times L})$$

A.2.2 MCMC Gibbs sampler

From the full Bayesian hierarchical model in *prior* section:

$$P(\mathcal{M}|D) \propto P(\mathcal{M})P(D|\mathcal{M})$$

$P(\mathcal{M}|D)$ is the posterior probability, $P(\mathcal{M})$ is the prior, and $P(D|\mathcal{M})$ is the likelihood. The above can be written out as:

$$\begin{aligned}P(\mathcal{M}|\mathbf{D})\propto & P(\mathbf{Y}|\mathbf{g}_{obs,k},\mathbf{g}_{mis,k},\mathbf{C},\mathbf{M},\Omega,\gamma^C,\gamma_k^M,\gamma_k^{\overline{M}},\sigma^2)\times P(\mathbf{g}_{obs,k}|\mathbf{M}_{obs},\Omega,\sigma_k^2)\\ &\times P(\mathbf{g}_{mis,k}|\mathbf{M}_{mis},\Omega,\sigma_k^2)\times P(\gamma_k^M|\mathbf{I}_k^M,\tau_k^M)\times P(\gamma_k^{\overline{M}}|\mathbf{I}_k^{\overline{M}},\tau_k^{\overline{M}})\times P(\Omega|\mathbf{I}_j^{\omega},\tau_j^{\omega})\\ &\times P(\mathbf{I}_k^M|\pi^M)\times P(\mathbf{I}_k^{\overline{M}}|\pi^{\overline{M}})\times P(\mathbf{I}_j^{\omega}|\pi^{\omega})\times P(\gamma^C)\times P((\tau_k^M)^2)\times P((\tau_k^{\overline{M}})^2)\\ &\times P((\tau_k^{\omega})^2)\times P(\pi^M)\times P(\pi^{\overline{M}})\times P(\pi^{\omega})\times P(\sigma^2)\times P(\sigma_k^2)\times P((\sigma^m)^2)\end{aligned}$$

We use MCMC Gibbs sampling as the approach to infer all posterior distributions. Some of the Gibb's samplers are listed below

- The effect of the j_0th methylation level on the mRNA level of kth gene:

$$[\omega_{j_0,k}|\cdot]\ \sim\ N(\sigma_n^2(A\sigma_k^{-2}+B(\gamma_k^M-\gamma_k^{\overline{M}})\sigma^{-2})'\mathbf{m}_{j_0},\ \sigma_n^2)$$

  where $A=(\mathbf{g}_k-M_{\mathcal{J}_k-j_0}\omega_{-j_0,k})$; $B=(\mathbf{Y}-\mathbf{C}\gamma^C-\mathbf{G}\gamma^{\overline{M}}+\mathbf{M}_{-j_0}\omega_{-j_0}(\gamma^{\overline{M}}-\gamma^M))$; $\sigma_n^2=(\sigma_{j_0}^{\omega\ -2}+(\sigma_k^{-2}+\sigma^{-2}(\gamma_k^M-\gamma_k^{\overline{M}})^2)\mathbf{m}_{j_0}'\mathbf{m}_{j_0})^{-1}$. For $k=1,...,K$, and $j_0\in\mathcal{J}_k$, which a the set of all $\mathbf{m}_{j_0}$ within the promoter of kth gene. $\mathbf{m}_{\mathcal{J}_k}=(\mathbf{m}_{j,j\in\mathcal{J}_k})$, and $\mathbf{m}_{-j_0}=(\mathbf{m}_{j,j\in\mathcal{J}_k,j\neq j_0})$; and similarly, $\omega_{\mathcal{J}_k,k}=(\omega_{j,k,j\in\mathcal{J}_k})$, $\omega_{-j_0,k}=\omega_{j,k,j\in\mathcal{J}_k,j\neq j_0}$
- The random error from the expression of gene k_0 that is NOT explained by methylation:

$$\left[\sigma_k^2|\cdot\right]\sim Inv.Gamma(\frac{N}{2}+\delta,\ \frac{1}{2}(\mathbf{g}_k^{\overline{M}})'(\mathbf{g}_k^{\overline{M}})+\delta)$$

  where this δ is a very small number originating from the Jeffery prior we put on σ_k^2
- The clinical effect:

$$\left[\gamma_l^C|\cdot\right]\sim N(\mathbf{C}_l'(\mathbf{Y}-\mathbf{M}\Omega\gamma^M-(\mathbf{G}-\mathbf{M}\Omega)\gamma^{\overline{M}}-\mathbf{C}\gamma_{-l}^C)(\sigma^2\sigma_0^{-2}+\mathbf{C}_l'\mathbf{C}_l)^{-1},(\sigma^{-2}\mathbf{C}_l'\mathbf{C}_l+10^{-6})^{-1})$$

- All Gibbs samplers for $\gamma's$ takes similar forms but different components. The effects of gene expression modulated by methylation, γ^M, is sampled from:

$$\left[\gamma_k^M|I_k^M=1,(\tau^M)^2\cdot\right]\sim N((\mathbf{M}_{\mathcal{J}_k}\omega_k)'A\sigma_n^2,\ \sigma_n^2)$$

  where $A=(\mathbf{Y}-\mathbf{M}_{-\mathcal{J}_k}\Omega_{-k}\gamma^M-(\mathbf{G}-\mathbf{M}\Omega)\gamma^{\overline{M}}-\mathbf{C}\gamma^C)$; and $\sigma_n^2=(\sigma^2(\tau^M)^{-2}+(M_{\mathcal{J}_k}\omega_k)'M_{\mathcal{J}_k}\omega_k)^{-1}$.

- The effects of gene expression *not* modulated by methylation, $\gamma^{\overline{M}}$, is sampled from:

$$\left[\gamma_k^{\overline{M}} | I_k^{\overline{M}} = 1, (\tau^{\overline{M}})^2 \cdot\right] \sim N(\mathbf{g}_k^{\overline{M}\prime} A \sigma_n^2,\ \sigma_n^2)$$

where $A = (\mathbf{Y} - \mathbf{M}\Omega\gamma^M - \mathbf{G}_{-k}^{\overline{M}}\gamma_{-k}^{\overline{M}} - \mathbf{C}\gamma^C)$; $\sigma_n^2 = (\sigma^2(\tau^{\overline{M}})^{-2} + \mathbf{g}_k^{\overline{M}\prime}\mathbf{g}_k^{\overline{M}})^{-1}$.

- The feature selection indicator for gene expression modulated by methylation I_k^M and its corresponding hyper prior π^M:

$$\left[I_k^M | \gamma_k^M, \pi^M\right] \sim Bernoulli(\frac{\pi_k^M f(\gamma_k^M | I_k^M = 1)}{f(\gamma_k^M | I_k^M = 0)(1 - \pi^M) + f(\gamma_k^M | I_k^M = 1)\pi^M})$$

$$\left[\pi^M | I_k^M\right] \sim Beta(1 + \sum_k I_k^M, 1 + K - \sum_k I_k^M)$$

- The feature selection indicator for gene expression *not* modulated by methylation $I_k^{\overline{M}}$ and its corresponding hyper prior $\pi^{\overline{M}}$:

$$\left[I_k^{\overline{M}} | \gamma_k^{\overline{M}}, \pi^{\overline{M}}\right] \sim Bernoulli(\frac{\pi_k^{\overline{M}} f(\gamma_k^{\overline{M}} | I_k^{\overline{M}} = 1)}{f(\gamma_k^{\overline{M}} | I_k^{\overline{M}} = 0)(1 - \pi^{\overline{M}}) + f(\gamma_k^{\overline{M}} | I_k^{\overline{M}} = 1)\pi^{\overline{M}}})$$

$$\left[\pi^{\overline{M}} | I_k^{\overline{M}}\right] \sim Beta(1 + \sum_k I_k^{\overline{M}}, 1 + K - \sum_k I_k^{\overline{M}})$$

- The feature selection indicator for effect of methylation j on gene k, I_k^ω and its corresponding hyper prior π^ω:

$$\left[I_j^\omega | \gamma_j^\omega, \pi^\omega\right] \sim Bernoulli(\frac{\pi_j^\omega f(\gamma_j^\omega | I_j^\omega = 1)}{f(\gamma_j^\omega | I_j^\omega = 0)(1 - \pi^\omega) + f(\gamma_j^\omega | I_j^\omega = 1)\pi^\omega})$$

$$\left[\pi^\omega | I_j^\omega\right] \sim Beta(1 + \sum_j I_j^\omega, 1 + J - \sum_j I_j^\omega)$$

- The error term in clinical model is:

$$\left[\sigma^2 | \cdot\right] \sim IG(\frac{N}{2} + \delta,\ \frac{1}{2}A'A + \delta)$$

where $A = \mathbf{Y} - (\mathbf{G} - \mathbf{M}\Omega)\gamma^{\overline{M}} - \mathbf{M}\Omega\gamma^M - \mathbf{C}\gamma^C$.

- The variance $(\tau^M)^2$, $(\tau^{\overline{M}})^2$, $(\tau^\omega)^2$ in spike-slab prior are sampled from:

$$\left[(\tau^M)^2|\cdot\right] \sim IG(\frac{1}{2}\sum_{k=1}^{K} I_k^M + \delta,\ \frac{1}{2}\sum_{k=1}^{K}(\gamma_k^M)^2 I_k^M + \delta)$$

$$\left[(\tau^{\overline{M}})^2|\cdot\right] \sim IG(\frac{1}{2}\sum_{k=1}^{K} I_k^{\overline{M}} + \delta,\ \frac{1}{2}\sum_{k=1}^{K}(\gamma_k^{\overline{M}})^2 I_k^{\overline{M}} + \delta)$$

$$\left[(\tau^\omega)^2|\cdot\right] \sim IG(\frac{1}{2}\sum_{j=1}^{J} I_j^\omega + \delta,\ \frac{1}{2}\sum_{j=1}^{J}(\gamma_j^\omega)^2 I_j^\omega + \delta)$$

Assuming missing in omics data, according to previous imputation model and corresponding prior structure, the posterior distributions of the missing omics data are:

- First, in missing gene expression imputation, for the initial MCMC iteration, for $b = 2, ...B$ iteration, the samples with missing gene expression are imputated as:

$$[\mathbf{g}_{mis,j}|\cdot] \ \sim\ MN_{N_{mis}^G}(\sigma_n^2(A\gamma_k^{\overline{M}}\sigma^{-2} + B\sigma_k^{-2}),\ \sigma_n^2)$$

where $\sigma_n^2 = (\sigma_k^{-2} + \sigma^{-2}(\gamma_k^{\overline{M}})^2)^{-1}$; $A = (\mathbf{Y}_{N_{mis}^G} - \mathbf{C}_{N_{mis}^G}\gamma^C - \mathbf{M}_{N_{mis}^G}\Omega(\gamma^M - \gamma^{\overline{M}}) - \mathbf{G}_{mis,-k}\gamma_{-k}^{\overline{M}})$; $B = (\mathbf{M}_{N_{mis}^G,\mathcal{J}_k}\ \omega_k)$. $\mathbf{G}_{mis,-k} = \mathbf{G}_{mis,i}\ where\ i \in (1...K), i \neq k$.

Then the complete methylation level is simply:

$$\mathbf{g}_j = (\mathbf{g}'_{obs,k}, \mathbf{g}'_{mis,k})'$$

- Then, in missing methylation imputation, for the initial MCMC iteration, for $b = 2, ...B$ iteration, the samples with missing methylation level are imputated as:

$$[\mathbf{m}_{mis,j}|\cdot] \ \sim\ MVN_{N_{mis}^M \times N_{mis}^M}(\sigma_n^2\omega_{j,k}(A(\gamma_k^M - \gamma_k^{\overline{M}})\sigma^{-2} + B\sigma_k^{-2}),\ \sigma_n^2\mathbf{I}_{N_{mis}^M \times N_{mis}^M})$$

where $\sigma_n^2 = (1+\omega_{j,k}^2(\sigma_k^{-2}+\sigma^{-2}(\gamma_k^M-\gamma_k^{\overline{M}})^2))^{-1}$; $A = (\mathbf{Y}_{N_{mis}^M} - \mathbf{C}_{N_{mis}^M}\gamma^C - \mathbf{M}_{mis,-j}\Omega_{-j}(\gamma^M - \gamma^{\overline{M}}) - \mathbf{G}_{N_{mis}^M}\gamma^{\overline{M}})$; $B = (\mathbf{g}_{N_{mis}^M,k} - \mathbf{m}_{mis,-j}\omega_{-j,k})$. And $\mathbf{m}_{mis,-j} = \mathbf{m}_{mis,i}\ where\ i \in \mathcal{J}_k, i \neq j$, $\mathbf{M}_{mis,-j} = \mathbf{M}_{mis,i}\ where\ i \in (1...J), i \neq j$, $\Omega_{-j} = \Omega_i\ where\ i \in (1...J), i \neq j$.

Then the complete methylation level is simply:

$$\mathbf{m}_j = (\mathbf{m}'_{obs,j}, \mathbf{m}'_{mis,j})'$$

Now that all the conditional probabilities are obtained, one may use Gibbs sampling to draw the posterior samples for the parameters following a certain sequence.

APPENDIX B

APPENDIX FOR LDIV IMPROVED GENOTYPE CALLING IN FAMLIY-BASED SEQUENCING DATA

B.1 SUPPLEMENTARY MATERIALS

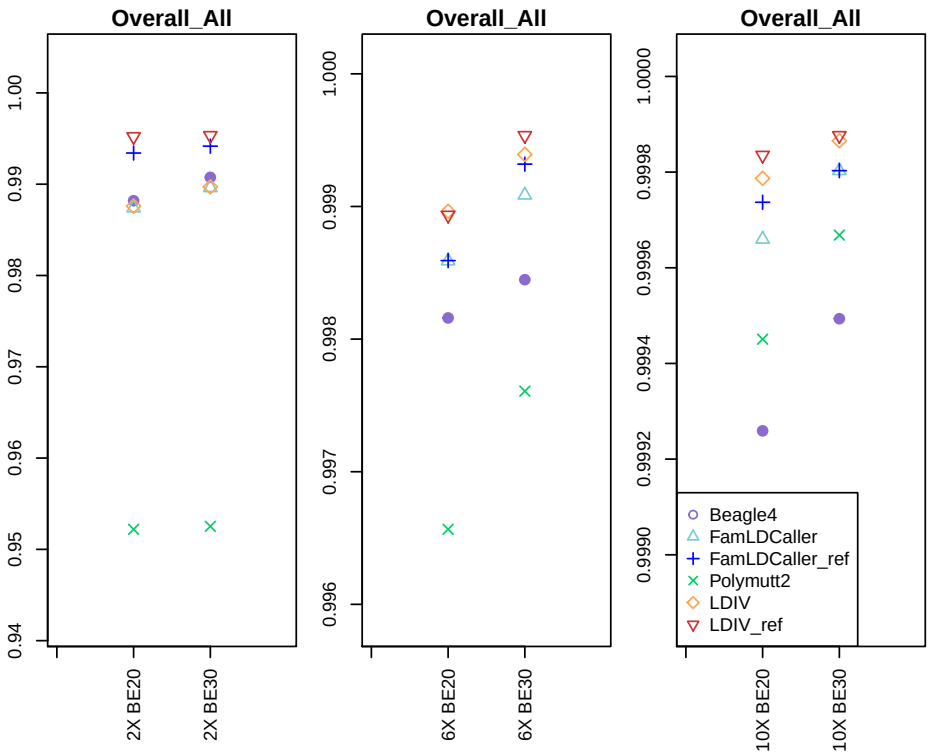


Figure 23: The overall genotyping accuracy for nuclear family type 3.

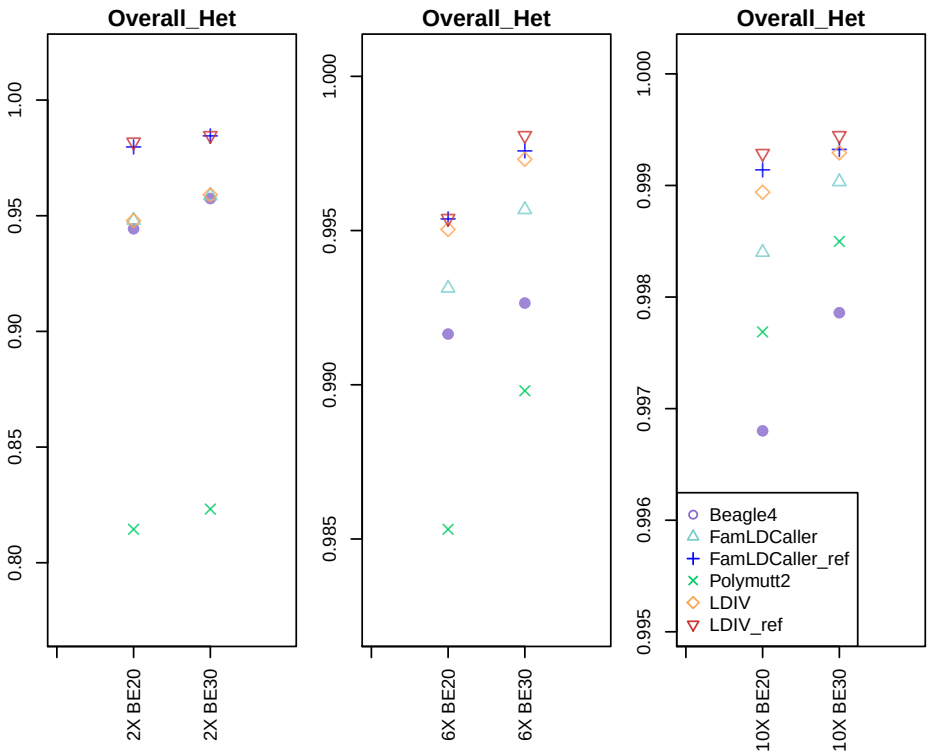


Figure 24: The genotyping accuracy on heterozygote sites for nuclear family type 3.

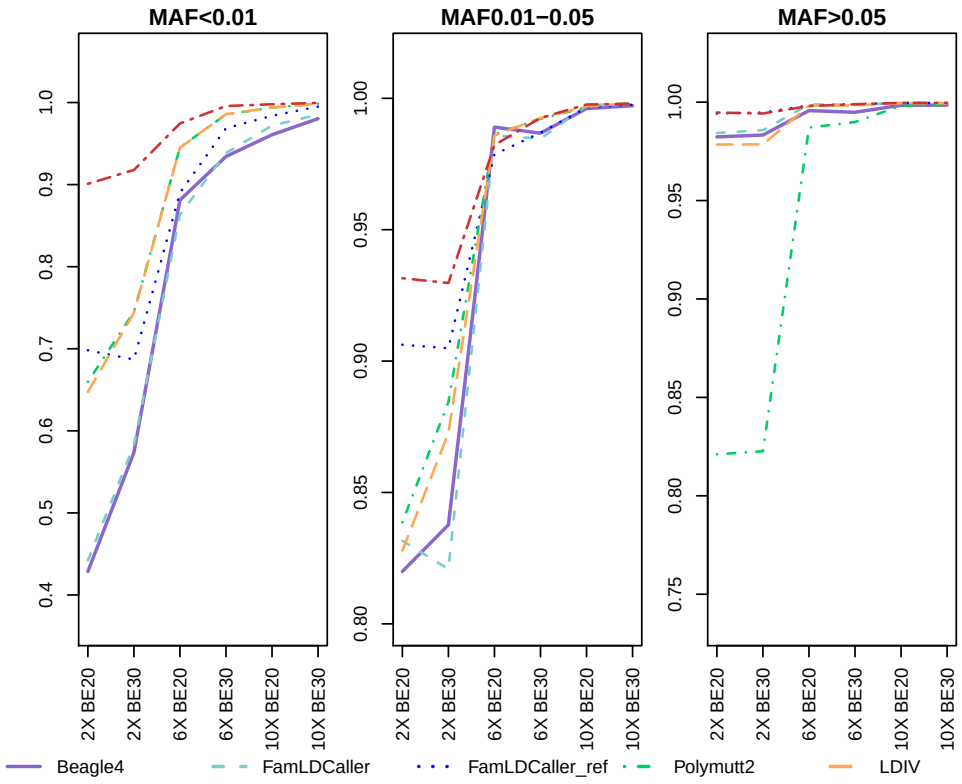


Figure 25: The genotyping accuracy on heterozygotes sites by MAF for nuclear family type 3.

Table 8: Overall genotyping accuracies for family type 3

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.9882	0.9874	0.9934	0.9522	0.9876	0.9952
2X BE30	0.9907	0.9896	0.9942	0.9525	0.9897	0.9953
6X BE20	0.9982	0.9986	0.9986	0.9966	0.999	0.9989
6X BE30	0.9984	0.9991	0.9993	0.9976	0.9994	0.9995
10X BE20	0.9993	0.9997	0.9997	0.9995	0.9998	0.9998
10X BE30	0.9995	0.9998	0.9998	0.9997	0.9999	0.9999

Table 9: Genotyping accuracies for heterozygote sites, family type 3

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.9444	0.948	0.9797	0.8145	0.9479	0.9819
2X BE30	0.9574	0.9586	0.9845	0.8232	0.9591	0.9847
6X BE20	0.9916	0.9931	0.9954	0.9853	0.995	0.9954
6X BE30	0.9926	0.9957	0.9976	0.9898	0.9973	0.9981
10X BE20	0.9968	0.9984	0.9991	0.9977	0.9989	0.9993
10X BE30	0.9979	0.999	0.9993	0.9985	0.9993	0.9994

Table 10: Genotyping accuracies for rare heterozygote sites, family type 3

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.4287	0.442	0.6981	0.6596	0.6474	0.9009
2X BE30	0.5735	0.5826	0.6869	0.7449	0.7437	0.9178
6X BE20	0.8812	0.8633	0.89	0.9452	0.9448	0.9745
6X BE30	0.9343	0.9386	0.9687	0.9854	0.9859	0.9957
10X BE20	0.9608	0.9719	0.9839	0.9939	0.9939	0.9979
10X BE30	0.9804	0.9851	0.9951	0.9985	0.9985	0.9996

Table 11: Genotyping accuracies for moderate heterozygote sites, family type 3

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.8199	0.8316	0.9062	0.8387	0.8279	0.9315
2X BE30	0.8377	0.8208	0.9049	0.8845	0.8726	0.9297
6X BE20	0.989	0.9859	0.9787	0.9862	0.9858	0.9822
6X BE30	0.9867	0.9846	0.9869	0.9922	0.9925	0.9925
10X BE20	0.9961	0.9973	0.9965	0.9966	0.9969	0.9976
10X BE30	0.9972	0.9983	0.9976	0.9982	0.9983	0.9979

Table 12: Genotyping accuracies for common heterozygote sites, family type 3

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.9824	0.9843	0.9942	0.8211	0.9785	0.9947
2X BE30	0.9833	0.9857	0.9947	0.8226	0.9786	0.9942
6X BE20	0.9957	0.9989	0.9983	0.9871	0.9982	0.998
6X BE30	0.9949	0.9987	0.999	0.9899	0.9984	0.9989
10X BE20	0.9984	0.9996	0.9996	0.9981	0.9995	0.9996
10X BE30	0.9985	0.9995	0.9996	0.9986	0.9995	0.9996

Table 13: Mendelian error rates for family type 3

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	28.82	14.22	8.833	0	13.33	7.689
2X BE30	25	16.36	7.189	0	9.222	6.989
6X BE20	6.422	3.667	2.289	0	0.7111	1.844
6X BE30	6.411	1.878	1.078	0	0.4222	0.8222
10X BE20	2.456	1.078	0.5444	0	0.2889	0.3222
10X BE30	2	0.7889	0.4111	0	0.2111	0.2778

Table 14: Phasing accuracy for family type 3

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.9882	0.9737	0.9917	0.8295	0.9591	0.9693
2X BE30	0.9907	0.9747	0.992	0.8373	0.9635	0.9662
6X BE20	0.9982	0.9964	0.9978	0.9615	0.9912	0.9712
6X BE30	0.9984	0.9959	0.9986	0.9551	0.9922	0.971
10X BE20	0.9993	0.9991	0.9994	0.9736	0.9972	0.9673
10X BE30	0.9995	0.9994	0.9994	0.9914	0.9992	0.9749

Table 15: Overall genotyping accuracies for family type 4

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.99	0.9869	0.9946	0.9622	0.9892	0.9959
2X BE30	0.992	0.9909	0.9958	0.9665	0.9909	0.9965
6X BE20	0.9985	0.9987	0.9992	0.9978	0.9993	0.9997
6X BE30	0.9987	0.9991	0.9994	0.9988	0.9995	0.9997
10X BE20	0.9994	0.9997	0.9998	0.9997	0.9999	0.9999
10X BE30	0.9995	0.9998	0.9998	0.9999	0.9999	0.9999

Table 16: Genotyping accuracies for heterozygote sites, family type 4

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.9515	0.9436	0.984	0.8586	0.9552	0.9843
2X BE30	0.9616	0.9594	0.9891	0.8773	0.9619	0.9879
6X BE20	0.9943	0.9944	0.9974	0.9922	0.9974	0.9987
6X BE30	0.994	0.9956	0.9977	0.9957	0.998	0.9989
10X BE20	0.9972	0.9987	0.9992	0.9987	0.9994	0.9996
10X BE30	0.998	0.9992	0.9992	0.9996	0.9996	0.9996

Table 17: Genotyping accuracies for rare heterozygote sites, family type 4

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.3675	0.3874	0.6135	0.6377	0.6376	0.8983
2X BE30	0.5886	0.5827	0.7833	0.8069	0.7877	0.9491
6X BE20	0.8763	0.8609	0.9517	0.9751	0.9751	0.9919
6X BE30	0.9489	0.9312	0.9625	0.984	0.984	0.9977
10X BE20	0.9509	0.9672	0.9737	0.9885	0.9885	0.9917
10X BE30	0.9725	0.9842	0.9839	0.9995	0.9995	0.9995

Table 18: Genotyping accuracies for moderate heterozygote sites, family type 4

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.8472	0.8165	0.9263	0.8751	0.8734	0.9293
2X BE30	0.8661	0.8627	0.9472	0.8976	0.8903	0.9467
6X BE20	0.9898	0.9834	0.986	0.9919	0.9921	0.9951
6X BE30	0.9902	0.9848	0.9871	0.9968	0.9967	0.9974
10X BE20	0.9959	0.9969	0.9974	0.9992	0.9992	0.9995
10X BE30	0.9972	0.9984	0.996	0.9995	0.9994	0.9991

Table 19: Genotyping accuracies for common heterozygote sites, family type 4

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.985	0.9814	0.9959	0.8654	0.9777	0.9955
2X BE30	0.9864	0.9851	0.996	0.8793	0.9793	0.9947
6X BE20	0.9967	0.9986	0.999	0.9927	0.9984	0.9993
6X BE30	0.9956	0.9987	0.9991	0.996	0.9985	0.9991
10X BE20	0.9986	0.9996	0.9997	0.999	0.9997	0.9998
10X BE30	0.9986	0.9997	0.9997	0.9996	0.9996	0.9996

Table 20: Overall genotyping accuracies for family type 5

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.9916	0.9891	0.9926	0.9546	0.9898	0.9947
2X BE30	0.9933	0.9904	0.9955	0.9591	0.9911	0.9967
6X BE20	0.9985	0.9988	0.9992	0.9947	0.9989	0.9995
6X BE30	0.9987	0.999	0.9992	0.9967	0.9993	0.9995
10X BE20	0.9994	0.9997	0.9997	0.9988	0.9997	0.9998
10X BE30	0.9995	0.9998	0.9998	0.9994	0.9998	0.9999

Table 21: Genotyping accuracies for heterozygote sites, family type 5

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.955	0.948	0.9747	0.7993	0.9494	0.9773
2X BE30	0.9632	0.9532	0.9843	0.8057	0.9555	0.9855
6X BE20	0.993	0.9939	0.9969	0.9753	0.9943	0.9974
6X BE30	0.9933	0.9946	0.9968	0.9818	0.996	0.9975
10X BE20	0.9974	0.9985	0.9989	0.9947	0.9986	0.9992
10X BE30	0.9978	0.9989	0.9992	0.997	0.999	0.9993

Table 22: Genotyping accuracies for rare heterozygote sites, family type 5

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.4024	0.414	0.5093	0.6381	0.6306	0.859
2X BE30	0.5763	0.5694	0.7177	0.7199	0.6971	0.9206
6X BE20	0.8761	0.8755	0.9004	0.9409	0.9409	0.9814
6X BE30	0.919	0.9128	0.9378	0.9604	0.9594	0.9863
10X BE20	0.9501	0.97	0.9815	0.9871	0.9871	0.9964
10X BE30	0.9743	0.983	0.984	0.993	0.9928	0.9958

Table 23: Genotyping accuracies for moderate heterozygote sites, family type 5

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.8627	0.8191	0.8476	0.8068	0.7809	0.8616
2X BE30	0.8879	0.843	0.9143	0.8559	0.8297	0.9182
6X BE20	0.9917	0.9826	0.986	0.9721	0.972	0.9877
6X BE30	0.9904	0.9864	0.9863	0.9872	0.9857	0.9892
10X BE20	0.9978	0.9961	0.995	0.9938	0.9943	0.9968
10X BE30	0.9981	0.9978	0.9973	0.9965	0.9966	0.9976

Table 24: Genotyping accuracies for common heterozygote sites, family type 5

	Beagle4	FamLDCaller	FamLDCaller_ref	Polymutt2	LDIV	LDIV_ref
2X BE20	0.9877	0.9839	0.9959	0.8055	0.983	0.9959
2X BE30	0.9873	0.981	0.9964	0.8038	0.9813	0.9967
6X BE20	0.9965	0.9987	0.9991	0.9769	0.9982	0.9991
6X BE30	0.9959	0.9985	0.9989	0.982	0.9984	0.999
10X BE20	0.9986	0.9996	0.9995	0.9951	0.9995	0.9995
10X BE30	0.9986	0.9996	0.9996	0.9973	0.9995	0.9997

BIBLIOGRAPHY

Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, 11(2):253–64.

Albrecht, D., Kniemeyer, O., Brakhage, A., and Guthke, R. (2010). Missing values in gel-based proteomics. *Proteomics*, 10(6):253–64.

Alexa, A. and Rahnenfuhrer (2010). topgo: Enrichment analysis for gene ontology. *R package version 2.18.0*.

Arion, D., Corradi, J., Tang, S., Datta, D., Boothe, F., He, A., Cacace, A., Zaczek, R., Albright, C., Tseng, G., and Lewis, D. (2015). Distinctive transcriptome alterations of prefrontal pyramidal neurons in schizophrenia and schizoaffective disorder. *Molecular psychiatry*, 20(11):1397–1405.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bird, A. (2002). Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21.

Bird, S. (2006). Nltk: the natural language toolkit. in proceedings of the coling/acl on interactive presentation sessions. *Association for Computational Linguistics*, pages 69–72.

Brock, G. N., Shaffer, J. R., Blakesley, R. E., Lotz, M. J., and Tseng, G. C. (2008). Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Biostatistics*, 9:1–12.

Browning, B. and Browning, S. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116 – 126.

Cancer Genome Atlas Network, . (2012). Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61.

Cancer Genome Atlas Research Network, . (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609.

Cancer Genome Atlas Research Network, . (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543.

Chang, L., Li, B., Fang, Z., Vrieze, S., McGue, M., Iacono, W., Tseng, G., and Chen, W. (2016). A computational method for genotype calling in family-based sequencing data. *BMC Bioinformatics*, 17:37.

Chen, W., Li, B., Zeng, Z., Sanna, S., Sidore, C., Busonero, F., and Abecasis, G. R. (2013). Genotype calling and haplotyping in parent-offspring trios. *Genome Research*, 23(1):142–151.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

DePristo, M. and Poplin, R. (2017). Deepvariant: Highly accurate genomes with deep neural networks. *Google Research Blog*, 1(2):3.

Edgar, R., Domrachev, M., and Lash, A. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Genomes Project Consortium, ., Abecasis, G., Auton, A., Brooks, L., DePristo, M., Durbin, R., Handsaker, R., Kang, H., Marth, G., and McVean, G. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–75.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gerstein, M., Lu, Z., and et al. (2010). Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, 330:1775–1787.

Geweke, J., Bernado, J., Berger, J., Dawid, A., and AFM, S. (1992). *Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bayesian Statistics 4*. Clarendon Press.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide

association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.

Huang, D. W., Sherman, B., and Lempicki, R. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocol*, 4(1):44–57.

Huo, Z. and Tseng, G. C. (2017a). Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *The Annals of Applied Statistics*, In press.

Huo, Z. and Tseng, G. C. (2017b). Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *Annals of Applied Statistics*, 11(2):1011–39.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Khatri, Purvesh, S. M. and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):1002375.

Kim, S., Oesterreich, S., Kim, S., Park, Y., and Tseng, G. C. (2017). Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics*, 18(1):165–179.

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC Hoen, P., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506.

Leinonen, R., Sugawara, H., Shumway, M., and on behalf of the International Nucleotide Sequence Database Collaboration, . (2011). The sequence read archive. *Nucleic Acids Research*, 39:D19–D21.

Lesk, A. (2017). *Introduction to genomics*. Oxford University Press.

Li, B., Wei, Q., Zhan, X., Zhong, X., Chen, W., Li, C., and Haines, J. (2015). Leveraging identity-by-descent for accurate genotype inference in family sequencing data. *PLOS Genetics*, 11(6):1–19.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., and Homer, N. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213â2233.

Li, Y., Sidore, C., Kang, M. H., Boehnke, M., and Abecasis, G. R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research*, 21(6):940–951.

Li, Y., Willer, C., Ding, J., Scheet, P., and Abecasis, G. (2010). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834.

Lin, D., Zhang, J., Li, J., Xu, C., Deng, H., and Wang, Y. (2016). An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics*, 17:247.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616.

McCarroll, S. A. and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics*, 39(7s):S37.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9(5):356.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., and DePristo, M. A. (2010). The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Metzker, M. L. (2010). Sequencing technologies–the next generation. *Nature reviews. Genetics*, 11(1):31.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118.

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 4:155–176.

Oh, S., Kang, D. D., Brock, G. N., and Tseng, G. C. (2011). Biological impact of missing-value imputation on downstream analyses of gene expression profiles. *Biostatistics*, 27(1):78–86.

Pan, D., Zhang, X., Huang, C., Jafari, N., Kibbe, W., Hou, L., and Lin, S. (2010). Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Reinhart, V. (2015). Evaluation of trkb and bdnf transcripts in prefrontal cortex, hippocampus, and striatum from subjects with schizophrenia, bipolar disorder, and major depressive disorder. *Neurobiology of disease*, 77:220–227.

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, 3(1):181–209.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 53-65:53–65.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York John Wiley & Sons.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147â177.

Schübeler, D. (2015). Function and information content of dna methylation. *Nature*, 517(7534):321.

Shen, K. and Tseng, G. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009a). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009b). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Subramanian, A. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Natl Acad. Sci. USA*, 102:15545–15550.

The 1000 Genomes Project Consortium, . (2015). A global reference for human genetic variation. *Nature*, 526:68–74.

The ENCODE Project Consortium, . (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, 40(9):3785–3799.

Tseng, G. C., Ghosh, D., and Zhou, X. J. (2015). *Integrating omics data*. Cambridge University Press.

Voillet, V., Besse, P., Liaubet, L., San Cristobal, M., and González, I. (2016). Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*, 17(1):402.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2012). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.

Watson, J. D. (1965). *Molecular biology of the gene*. New York: W. A. Benjamin.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. M., Ozenberger, B. A., Ellrott, K., and et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Zhou, B. and Whittemore, A. S. (2012). Improving sequencing-based genotype calls with linkage disequilibrium and pedigree information. *The Annals of Applied Statistics*, 6(2):457–475.